



**TEXT TO SPEECH FOR KAMBAATA LANGUAGE BY USING
STATISTICAL PARAMETRIC APPROACH**

MSc THESIS

BY

ABENEZER SIYOUM GANORO

JUNE 2023

HOSSANA, ETHIOPIA

TEXT TO SPEECH FOR KAMBAATA LANGUAGE BY USING
STATISTICAL PARAMETRIC APPROACH

BY

ABENEZER SIYOUM GANORO

ADVISOR: ABDULLAH MOHAN (PHD)

A THESIS SUBMITTED TO THE DEPARTMENT OF ENGINEERING
COLLAGE OF ELECTRICAL AND COMPUTER ENGINEERING,
SCHOOL OF GRADUATE STUDIES
WACHEMO UNIVERSITY
HOSSANA, ETHIOPIA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE IN **ELECTRICAL AND
COMPUTER ENGINEERING**
(SPECIALIZATION: **COMPUTER ENGINEERING**)

JUNE 2023
HOSSANA, ETHIOPIA

**SCHOOL OF GRADUATE STUDIES
WACHEMO UNIVERSITY**

ADVISORS' APPROVAL SHEET

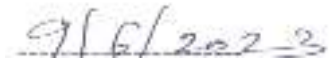
We, the undersigned, members of the Board of Examiners of the final open defense have read and evaluated her thesis entitled **“Text To Speech For Kambaata Language By Using Statistical Parametric Approach”**, and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirements for the degree.

Abdullah Mohan (PhD)

Name of major advisor



Signature



Date

Name of co-advisor

Signature

Date

EXAMINERS' APPROVAL SHEET

_____	_____	_____
Name of the Chairperson	Signature	Date

Name of Major Advisor

Signature

Date _____

Name of Internal Examiner

Signature

Date _____

Almaw Ayele Aniley (PhD)



08/07/2023G.C

Name of External examiner

Signature

Date _____

SGS Approval

Signature

Date _____

Declaration

I hereby declare that this MSc thesis entitled “Text to Speech for Kambaata Language by Using Statistical Parametric Approach” contains fully original research work carried out by me under the supervision of Dr. Abdullah Mohan. This thesis has been submitted in partial fulfillment for the requirement of the degree of Master's with specialization in Computer engineering, to Wachemo University. I earnestly declare that this thesis has not been presented for the award of any other degree, diploma, fellowship, or other similar titles or prizes in any other university; and all sources of material used for this thesis have been duly acknowledged.

Name of the student: **Abenezer Siyoum Ganoro**

Place: Hossana

Date: 14/07/2023G.C

Signature:



Acknowledgement

My first and foremost acknowledgement goes to God, who has helped me not only in doing this research but also in any step of my life.

I am indebted to my advisor, Dr. Abdullah Mohan, for his patience and guidance throughout the research period. His guidance helped me through all the research and writing of this thesis.

My heartfelt thanks go to my parents for their unreserved support in every way of my life and in each step of my educational endeavor. A special thanks to my father, Siyoum Ganoro, and beloved mother, Martha Sugamo, for standing by my side all the way through my study.

It is my pleasure to acknowledge all the respondents who participated in the questionnaire also those who were involved directly or indirectly in the completion of this study, including members of the Kambaata Tembaro Zone Culture and Tourism Office and others.

My endless gratitude and deepest appreciation go to my friend, who assisted me during this research study. I thank you a lot and appreciate your help. Your diligent efforts have been decisive for the quality and effective accomplishment of this study, and you will always be in the conscious part of my mind.

Abstract

Speech synthesis is the process of creating human speech inanimately and can be used in hardware or software applications. Today's modern human activities, including assistance for the disabled and the telecommunications industry, involve speech synthesis in many ways. Its task is to convert written text into speech. The most important criteria for evaluating a synthesised speech's quality are its naturalness and intelligibility. The two primary components of text-to-speech, which are usually referred to as the frontend and backend systems, are natural language processing and digital processing. In this study, the first text-to-speech system for Kambaata is demonstrated. For this research, statistical parametric voice synthesis based on HMM approaches was chosen because it is model-based, requires minimal storage, learns data attributes rather than storing speech, has a short run time, and is simple to integrate with small mobile devices. Several steps made up of functional components are used to convert the input text into an acoustic waveform. The training and testing phases are the two primary parts of the synthesizer. For the speech that is taken out of speech databases during the training phase, EHMM labeling is used to automatically segment the speech and produce phonetic transcriptions. In the testing phase, the trained models are combined with the input text to create phonetic strings. Finally, speech parameters are used to create the synthesised speech. I gathered 400 sentences and speeches to train the system that was being created. I also tested the system's effectiveness using five sentences. This study used the Mean Opinion Score (MOS) evaluation method. The system's efficiency is tested, and the correctness and naturalness of the synthesised speech are evaluated. Using MOS testing methods, the intelligibility and naturalness of synthesised speech are evaluated. The results yield scores of 2.97 and 2.5, respectively.

Keywords: Statistical Parameter Speech Synthesis, Text to Speech, Kambaata, and Hidden Markov Model based speech synthesis.

Contents

Declaration	ii
Acknowledgement.....	iii
Abstract	iv
Contents.....	vi
List of table.....	ix
List of Figure	x
List of Abbreviations.....	xi
Chapter One	1
Introduction.....	1
1.1 Background	1
1.2 Statement of the Problem	3
1.3 General Objective.....	4
1.4 Specific Objectives.....	4
1.5 Scope of the study	4
1.6 Research gap	4
1.7 Research Methodology.....	5
1.7.1 Literature Review	5
1.7.2 Data Collection	5
1.7.3 Modeling.....	5
1.8 Testing and Evaluations Techniques.....	6
1.9 Significance of the study	6
Chapter Two.....	7
Literature review	7
2.1 Introduction	7
2.2 Historical background of Speech Synthesis	7
2.3 Human Speech Production Systems.....	9

2.3.1 Anatomy	9
2.3.2 Physiology and Function	10
2. 4 Source Filter Model.....	11
2.5 Basics of Text-to-Speech (TTS) Synthesis	14
2.6 Natural Language Processing (NLP).....	15
2.6.1 Text Analysis.....	16
2.6.2 Phonetic Analysis	16
2.6.3 Prosodic Analysis	17
2.7 Digital Signal Processing (DSP)	18
2.8 Speech synthesis techniques.....	18
2.8.1 Articulatory Synthesis	18
2.8.2 Formant Synthesis	19
2.8.3 Concatenative Synthesis	19
2.8.4 HMM Based Synthesis	20
2.9 Speech Synthesis System	20
2.9.1 Festival Speech Synthesis System.....	20
2.9.2 Festvox Speech Synthesis System.....	21
2.10 Related Works	21
2.10.1 Text to Speech Researches on International Languages	21
2.10.2 Text to speech Researches on Local Languages	22
2.11 Kambaata language and Phonology	25
2.11.1The Kambaata People and the language.....	25
2.11.2 Phonology.....	27
Chapter Three.....	42
Methodology and Materials	42
3.1 The Text to Speech Algorithm.....	42
3.2 Overview of statistical parametric synthesis.....	42

3.3 The Hidden Markov Model.....	43
3.4 Mel-Cepstrum Analysis	44
3.5 Linear Prediction Model.....	44
3.6 Fundamental Frequency Modeling.....	44
3.7 HMM state duration Modeling.....	45
3.8 Linear Prediction Model.....	45
3.9 CLUSTERGEN synthesizer.....	45
3.10 Decision Tree Building for Context Clustering	46
3.11 Text to speech synthesis system architecture.....	48
3.11.1 Introduction	48
3.11.2. Training Phase	49
Chapter Four	53
Results and Discussion	53
4.1 The Development Environment and Tools	53
4.2 Preparing Questionnaire.....	53
4.3 Testing and Evaluations	54
Chapter Five.....	56
Conclusions and Recommendation.....	56
5.1 Conclusions	56
5.2 Recommendations	57
Reference.....	58
Appendix	62

List of table

Table 2.1 some related work in the form of table	25
Table 2.2 Consonant phonemes	35
Table 2.3 the occurrence of vowels in word-initial position	35
Table 2.4 the occurrence of vowels in word-medial position.....	36
Table 2.5 the occurrence of vowels in word-final position	36
Table 4.1 Scales used in MOS	54
Table 4.2 MOS score of the sound.....	55

List of Figure

Figure 2.1 The VODER speech synthesizer [22].....	8
Figure 2.2 An overview diagram of the major articulators [27]	11
Figure 2. 3 Block diagram of the human speech production system [28]	12
Figure 2.4 Block diagram of a basic voicing source model [29]	13
Figure 2.5 Stephen Hawking [32].....	14
Figure 2.6 Basic Architecture of a TTS system [35]	15
Figure 2.7 Factors contributing to prosodic feature [2]	17
Figure 2.8 Kembata tembaro administrative map [14]	26
Figure 2.9 The branch of Cushitic family [14]	27
Figure 3.1 left-right Markov models [49].....	43
Figure 3.2 Decision Tree Building for Context Clustering [52].....	47
Figure 3.3 HMM-based speech synthesis system [53]	48
Figure 3.4 Proposed architecture of speech synthesizer for Kembatta language by using statistical parametric approach.....	49
Figure 4.1 MOS Score of the Sound in graph.....	55

List of Abbreviations

ASR	Automatic Speech Recognition
CART	Classification and Regression Tree
DSP	Digital Signal Processing
LPC	Linear predictive Code
LTS	Letter to Sound
EHMM	Ergodic hidden Markov models
HCI	Human Computer Interaction
HMM	Hidden Markov Model
HRL	High Resource Language
HTK	Hidden Markov Model Tool Kit
HTS	HMM-based speech synthesis system
IPA	International Phonetic Alphabet
MCD	Mel Cepstral Distortion
MFCC	Mel frequency cepstral Coefficients
NLP	Natural Language Processing
NSW	Non Standard Words
SGD	Speech Generating Device
SPSS	Statistical Parametric Speech Synthesis
SPTK	Speech Processing Tool Kit
TTS	Text To speech
UTS	Unit Selection Technique

Chapter One

Introduction

1.1 Background

One of the best ways for humans to communicate is through speech. It is especially important in interactions between humans and machines. We now have a wide variety of tools in this field because of technological improvements. Because they provide hands-free, natural, and universal access to the interacting device, speech-enabled interfaces are appealing. For many years, speech synthesis the automated production of speech waveforms—has been under research. Although recent advancements in voice synthesis have led to synthesisers with extremely high intelligibility, sound quality and naturalness are still significant issues. In addition, the majority of speech synthesisers now lack versatility since they only have a limited number of parameters they can work with. However, the quality of the items on the market today has improved enough for several applications, such as multimedia and telecommunications. [1] [2]

Text-to-speech synthesis (TTS) is one of the key technologies in speech processing techniques for creating speech signal from arbitrarily given text in order to transmit information from a machine to a person by voice [3]. The main goal of Text-to-speech synthesis (TTS) synthesis is to produce a natural and intelligible sounding speech from arbitrary text [4]. Moreover, the current trend in Text-to-speech synthesis (TTS) research calls for systems that enable producing speech in different speaking styles with different voice characteristics with small run time and emotions. [5] In order to fulfill these requirements, the best suitable approach is statistical parametric speech (SPS) synthesis system technique based on hidden Markov model (HMMs) [6].

Over the past many years, speech technology has made significant advancements [7] . The concept that a machine could produce speech was realized, but such devices weren't actually viable until the last 50 years. For instance, the U.K. Telephone Company created a speaking clock in 1936, which was one of the first real-world uses of voice synthesis. [7] The first fully functional text-to-speech system was finished in 1968, and the first computer-based speech synthesis systems were developed in the late 1950s. [8]

Naturalness and Intelligibility are two of a speech synthesis system's key characteristics [8] [9] [10]. Naturalness refers to how closely the output resembles human speech, whereas intelligibility refers to how simple it is to comprehend the output. A voice synthesiser should sound both natural and understandable. Typically, speech synthesis systems aim to maximize both qualities. [8].

The TTS system is divided into two phases the study of how computers and natural (congenital) languages interact is known as natural language processing, or NLP. NLP is a subfield of computer wisdom, artificial intelligence, and computational linguistics. NLP is associated with the field of earthborn- computer commerce as a result. The alternate stage is the Digital Signal Processing (DSP) step, which is in charge of voice creation procedures. NLP is significant in machine restatement, fighting spam, information birth, summarization, question answering, speech- recognition, and speech conflation, among other effects. [11] [12]The DSP module's conditioning may be allowed of as the computer fellow of stoutly regulating the vibratory frequency of the oral crowds and the articulatory muscles similar that the affair signal fits the input conditions. [10]

The syntactic structure and semantic focus of the sentence are used by the text analysis procedure to translate the input text into abstract linguistic description (such as phonemes and stress). The "Text Normalization" approach, which stretches concept-bearing character sequences into language-specific word sequences, is used to process input texts first. The voice creation procedure initially creates the phonetic realization of each phoneme using the linguistic framework. The phonetic-to-acoustic conversion, also known as speech synthesis, is then carried out. Based on the speech parameters, the phonetic-to-acoustic transformation executes the synthesis. Typically, these two stages are referred to as high-level and low-level synthesis. [7] [13]

In Ethiopia speech synthesis has been developed for different languages and it is a demanding research area for other languages as well. Up to now, Kambaata is a language which has no speech synthesis developed so far among others.

Kambaata(Self- appellation = Kambaatissata) is a language of the Kambaata- Xambaaro Zone of the Southern Nations, Ethnicities, and Peoples Region(SNNPR) of Ethiopia and by Kambaata settlers in other corridor of the country. Kambaata speaker number 685, 167, is according to the 2007 Ethiopian population and population report, with 337, 852 men and

345, 315 women. They live in the upland areas around the Abarrichcho massif about 300 km southwest of the Ethiopian capital, Addis Ababa, between the Omo River to the west and the Billate River to the east. [14] The KambaataTembaaro Zone comprises three major ethnical groups Kambaata, Tembaaro and Donga. The capital of the zone is Duuraame, 300 km from Addis Ababa Kambaata uses the Latin- grounded rudiments and it's a sanctioned language of Kambaata- Xambaaro Zone and also it has been used as medium of instruction for primary academy, as language class at inferior and secondary seminaries and language study at Duuraame preceptors Training College and Wachamo University. Currently, literatures, books, journals publishing in Kambaata have increase over the times.

1.2 Statement of the Problem

The importance of Natural language processing applications such as machine translation, speech-recognition, part-of-speech tagging, speech synthesis etc. becomes increasing for our local languages. Among these applications, speech synthesis is a useful tool to contribute a lot in the effort of NLP applications development. It is playing different roles in human day to day activities like aid for the disabled people in their activities and used in different commercial areas in telecommunication system and also in academic area as knowledge management. In fact there are speech synthesis that has been developed for foreign languages and local languages. Presently, there are many functional speech synthesizers which are developed and used in foreign languages such as Japan, Vietnam, Bengali and English [15] [16] [17], Ethiopian researchers have been able to develop a prototype TTS synthesizer model for regional languages such as Amharic, Afaan Oromo, wolaytta and Tigrigna [18] [19] [8] [20]. However, these speech synthesis cannot be applied directly for Kambaata, Language; Due to the differences in the internal structure of the words, syntactic (structure of the sentence) and semantic (meaning of the word) of a languages. Therefore; developing speech synthesis for Kambaata, Language contributes a lot in the field of Natural Language Processing and puts a benchmark for other higher level NLP applications development for the future.

The following research questions are formulated to pursue the research work:

- Which model of speech synthesis is being used for best synthesis for Kambaata, language?
- How to implement statistical parametric speech synthesis for Kambaata, language?

1.3 General Objective

The general objective of this study is to design speech synthesizer for Kambaata language by using statistical parametric approach.

1.4 Specific Objectives

- ❖ To prepare Kambaata corpus.
- ❖ To design and implement a Text to Speech (TTS) for synthesis for Kambaata language by using HMM techniques.
- ❖ To undertake experimentation to measure the performance of the system on selected test dataset.
- ❖ To test the performance and draw conclusions and recommendations based on the experiment results.

1.5 Scope of the study

This research deals with speech synthesizer for Kambaata language by using Statistical parametric approach. However nonstandard words such as, time, acronym, abbreviation, date and other numeric are not considered. But these will be considered as future work.

1.6 Research gap

The Text to Speech technology is beneficial for the in school visually impaired student's language acquisition as well as reading and communication needs. In addition to the previously mentioned data, researchers employ various methodologies in accordance with various speech synthesis techniques, such as diphone concatenative synthesis, formant synthesis, and using unit selection method, in order to develop a prototype for a Text to Speech system for various local and international languages. These approaches are of poor quality and feature enormous, meticulously labeled data bases. These studies know of no Text to Speech (TTS) systems that have been built for the Kambaata language. Therefore, the purpose of this research effort is to develop a text-to-speech synthesis for the Kambaata language using an advanced (Statistical parametric speech synthesis) technique that produces comprehensible speech, which is essential for many application areas.

1.7 Research Methodology

1.7.1 Literature Review

To understand Natural Language Processing and Kambaata language (phonological property of the language) deeply related literature have been reviewed and especial focus, in the area of speech synthesis (model-based speech synthesis) given.

1.7.2 Data Collection

There was no readymade text and speech corpus for Kambaata that used for this work. So classes were assessed and identified from different literatures and consulted experts. And also raw texts collected for corpus development from different datasets of Kambaata student books, translated newsletters and University Kambaata language teaching modules and sentences be record with experienced experts.

1.7.3 Modeling

This thesis used model-based speech synthesis prototype for Kambaata language idiom. Statistical parametric speech synthesis based on HMM techniques was chosen for these research because it is a model based that require less storage, it learn properties of data rather store the speech, small run time, and easy to integrate with small handheld devices.

1.7.4 Development Tools

A statistical parametric technique has been used to construct a Text to Speech (TTS) for the Kambaata language. Festival speech synthesis is the name of the toolkit that is used by this speech synthesis system. The Festival speech synthesis system is a multi-platform, general multi-lingual speech synthesis work branch. The method that is most widely used now is statistical parametric speech synthesis, which is versatile because it is model-based speech synthesis. The HMM-based speech synthesis system's toolkit is known as HTS. The creation of text-to-speech systems applying Markov models is the focus of the HMM-based Speech Synthesis System, a set of open source tools. These tools' content is limited to voice production, modeling, and training with no text processing. Additionally, several software toolkits from recording to synthesis are employed with the Windows operating system. Ubuntu LTS can also be used to run software. Adobe Audition 12.1.5.3 is used as the recording tool, along with a wired microphone and earphones for voice recording.

1.8 Testing and Evaluations Techniques

The overall TTS system is assessed based on intelligence and naturalness. The intelligence of a system is how much of the spoken output the user understands, as well as how quickly a listener gets fatigued by only listening, whereas the naturalness of a system is the measure of how much like real speech the output of the TTS system sounds. The mean opinion score (MOS) is the most commonly used technique to evaluate the overall performance of any TTS system, even its naturalness. MOS uses a scale level that ranges from 1 (bad) to 5 (excellent) as per the system's performance. For this specific research work, the Mean Opinion Score (MOS) scale is used in preference to the others because it is widely used and is the simplest method to evaluate the overall performance of a speech quality.

1.9 Significance of the study

The study can lay the groundwork for further research in the areas of speech synthesis, in particular in the Kambaata language. It will help economically in the telecommunication fields, in cultural and social affairs for disabled people, and academically in knowledge management as a policy and curriculum source for learning Kambaata.

Chapter Two

Literature review

2.1 Introduction

Speech is the key to communication between human beings. This section discusses the historical background of speech synthesis, the human speech system basics of text-to-speech (TTS) synthesis, the fundamentals of speech synthesis, and speech synthesis techniques. The speech synthesis phases are discussed briefly, with a focus on statistical parametric speech synthesis. Finally, I presented a review of related works on text-to-speech research in international languages and in local languages.

2.2 Historical background of Speech Synthesis

The history of synthesized speech from the first mechanical efforts to today's high-quality synthesizers took a long period of time. Speech synthesis mainly has gone through two stage of development; mechanical and electronic stage of development [7]. In 1779 Russian Professor Christian Kratzenstein explained physiological differences between five long vowels (/a/, /e/, /i/, /o/, and /u/) and made apparatus to produce them artificially [21]. He constructed acoustic resonators similar to the human vocal tract and activated the resonators with vibrating reeds like in music instruments.

A few years later Wolfgang von Kempelen introduced his "Acoustic-Mechanical Speech Machine" which was able to produce single sounds and some sound combinations [22]. Kempelen also published a book in which he described his studies on human speech production and the experiments with his speaking machine. The essential parts of the machine were a pressure chamber for the lungs, a vibrating reed to act as vocal cords, and a leather tube for the vocal tract action. His studies led to the theory that the vocal tract, a cavity between the vocal cords and the lips, is the main site of acoustic articulation. [22]

In late 1800's Alexander Graham Bell with his father, inspired by Wheatstone's speaking machine, constructed same kind of speaking machine [7] [21]. The research and experiments with mechanical and semi-electrical analogs of vocal system were made until 1960's, but with no remarkable success.

The first full electrical synthesis device was introduced by Stewart in 1922. [23] The synthesizer had a buzzer as excitation and two resonant circuits to model the acoustic resonances of the vocal tract. The machine was able to generate single static vowel sounds with two lowest formants, but not any consonants or connected utterances. [23]

First device to be considered as a speech synthesizer was VODER (Voice Operating Demonstrator) introduced by Homer Dudley in New York World's Fair 1939. [22] VODER was inspired by VOCODER (Voice Coder) developed at Bell Laboratories in the mid-thirties. The original VOCODER was a device for analyzing speech into slowly varying acoustic parameters that could then drive a synthesizer to reconstruct the approximation of the original speech signal. [22]

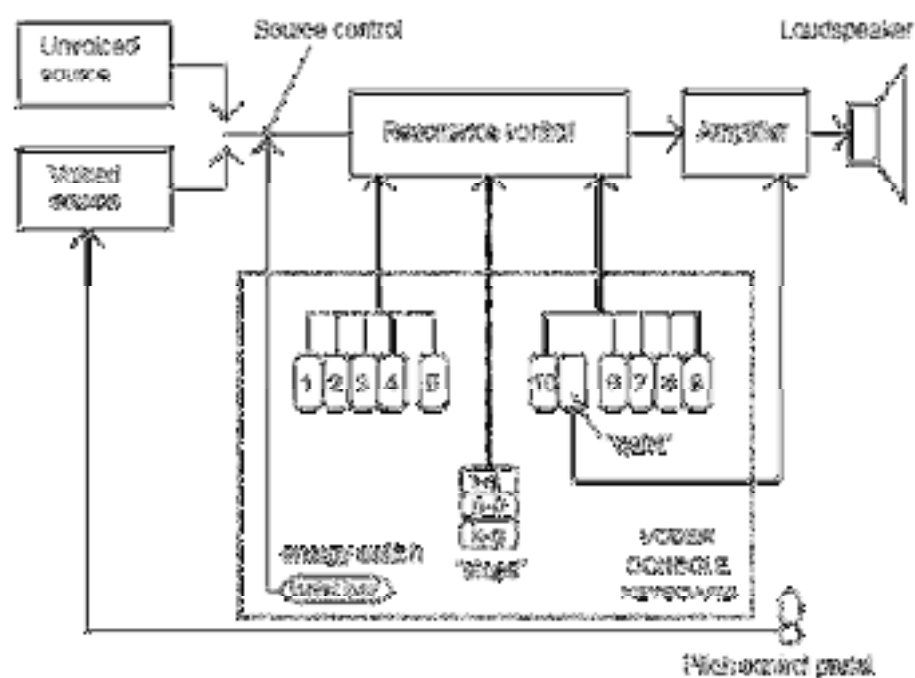


Figure 2.1 The VODER speech synthesizer [22]

After demonstration of VODER the scientific world became more and more interested in speech synthesis. It was finally shown that intelligible speech can be produced artificially. The first formant synthesizer, PAT (Parametric Artificial Talker), was introduced by Walter Lawrence in 1953. [7] PAT consisted of three electronic formant resonators connected in parallel. First articulatory synthesizer was introduced in 1958 by George Rosen at the Massachusetts Institute of Technology, M.I.T. [7]

Modern speech synthesis technologies involve quite complicated and sophisticated methods and algorithms. Methods applied recently in speech synthesis are hidden Markov models (HMM) and neural networks. These speech syntheses systems have been used for about decades and shows latest results have been quite promising.

2.3 Human Speech Production Systems

The mechanisms of Speech production can be divided into 3 systems: the air pressure (breathing mechanism), vibratory (the vocal folds), and the resonating systems (the supra glottic airway and vocal tract)

2.3.1 Anatomy

- I. The breathing mechanism or power source includes the lungs, diaphragm and chest wall muscles
- II. The vibratory system or sound source includes the vocal folds
 - The intrinsic laryngeal muscles control the shape of the glottis and the length and tension of the vocal folds; all of these muscles are innervated by the terminal branch of the recurrent laryngeal nerve (RLN), except for the cricothyroid muscle which is innervated by the external branch of the superior laryngeal nerve (SLN).
 - The intrinsic laryngeal muscles include 3 adductors, 1 abductor and 1 tensor:
 - ✓ **Adductors:** thyroary tenoid (TA), lateral cricoary tenoid (LCA), and interary tenoid (IA), which also consists of transverse and obliqueary tenoid fibers.
 - ✓ **abductor:** posterior cricoary tenoid (PCA)
 - ✓ **tensor:** crico thyroid (CT)
 - the vocal folds are composed of three layers: mucosa, vocal ligament and muscle
 - ✓ The **mucosa** includes the squamous epithelium, basement membrane, and the superficial lamina propria (Reinke's space)
 - ✓ The **vocal ligament** includes the intermediate and deep layers of the lamina propria (LP) layer

- ✓ The **muscle** includes the thyroaryv tenoid/vocal is muscle
 - The anterior 2/3 of the vocal fold in adults is typically the membranous orphonatory portion, while the posterior 1/3 is the cartilaginous or respiratory portion.
 - The vascular supply to the larynx comes from the superior and inferior larynx gealarteries and veins.
- III. The resonating system or vocal tract include the supra glottic larynx, pharynx, oral cavity, and nasal cavities. [24]

2.3.2 Physiology and Function

- I. The process of voice production is outlined below:
 - Glottic closure -----during exhalation facilitates increase in subglottic pressure ---- once subglottic pressure exceeds glottic closure force ---- air passes through the vocal folds ---- mucosal wave begins infra glottically and is propagate d super laterally glottic pressure drops due to open phase + elastic recoil of tissues ---- leads to glottal closure ---- which facilitates increase in subglottic pressure & glottal cycle repeats
- II. The primary motor neurons controlling the intrinsic laryngeal muscles are located in the nucleus ambigu us, which receives both excitatory and inhibitory input from the brainstem, controlling respiration, cough, and swallowing. For phonation (volitional movement), there is direct innervation from the cerebral cortex. For emotional vocalizations, there are additional connections from the limbic system.
- III. The microanatomy of the vocal folds is specialized such that the pliable cover layer can vibrate freely over the stiffer body under layer, creating a mucosal wave.
- IV. The vocal tract individualizes the human voice by acting as both a resonator and filter of the sound created by the vocal folds,
 - Vocal fold vibration results in a fundamental frequency (F0) of vibration, as well as many other frequencies or Harmonics, whose frequencies are integer multiples of the fundamental frequency.
 - Formants are a result of vocal tract shape and refer to how the vocal tract changes the relative amplitude of the harmonic spectrum. The vocal tract can actively change shape (e.g. different shapes for different vowels), which leads

to unique shaping of the amplitudes of harmonics into unique format structures.

- Resonance is achieved when the harmonics and formants are the same, resulting in an increased intensity of the sound created; a resonant singing tone is desirable in classical music and allows a singer to be easily heard over an orchestra without amplification. [25] [26] [27]

The exact placement of the main organs is shown in figure 2.2

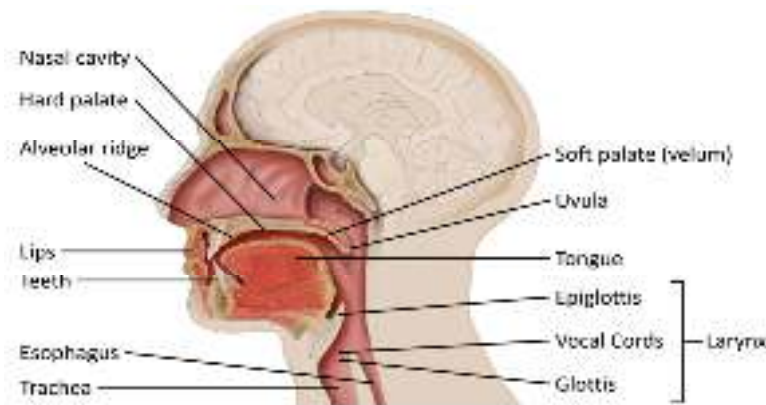


Figure 2.2 An overview diagram of the major articulators [27]

2. 4 Source Filter Model

Nearly all techniques for speech synthesis and recognition are based on the model of human speech production system. The central idea here is the decomposition of the speech signal as a source passed through a linear time-varying filter. This filter can be derived from models of speech production based on the theory of acoustics that states the source represents the air flow at the vocal cords and the filter represents the resonances of the vocal tract, which change over time. Such a source-filter model is illustrated in Figure 2.3 [28] The filter (i.e. a set of resonators) is excited by a source, which can be either a simulation of vocal cord vibration for voicing, or a noise that simulates a constriction somewhere in the vocal tract. The sound wave is created in the vocal tract, and then radiates through the lips. [29]

In this model, there is no interaction between the source and the filter other than the fact that the filter imposes its resonant characteristics on the source. [29] Hence, the individual acoustic properties of the source and the filter can be separately simulated. The vocal tract filter can be modeled as an acoustic tube with a varying cross-sectional area formed by the pharynx, the oral cavity, the nasal cavity, and the lips.

The speech sounds generated may group into two broad categories. Those produced with a periodic vibration of the vocal cords (voiced sounds), and those generated without vocal-cord vibrations, but with plosive or friction noise (voiceless sounds). For this reason, two excitation sources are needed for synthesis: a source producing a quasi-periodic wave (the voicing source) and a noise generator (the friction source). [29]

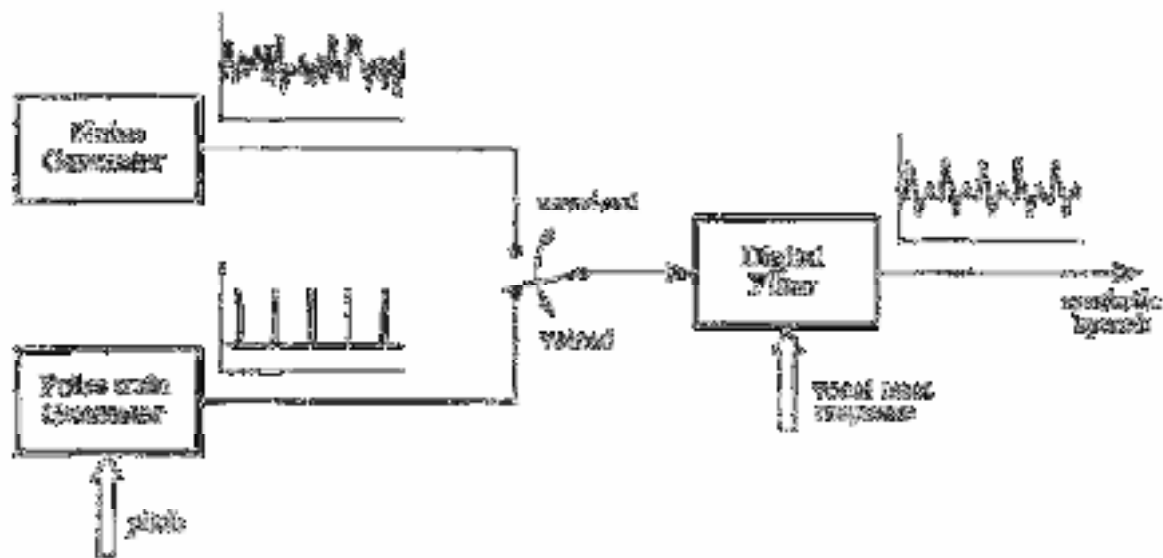


Figure 2. 1 Block diagram of the human speech production system [28]

Voiced sounds occur when air is forced from the lungs, through the vocal cords, and out of the mouth and/or noise. [8] [28] [29] The vocal cords are two thin flaps of tissue stretched across the air flow, just behind the Adam's apple. In response to varying muscle tension, the vocal cords vibrate at frequencies between 50 and 1000 Hz, resulting in periodic puffs of air being injected into the throat. The rate of cycling (opening and closing) of the vocal folds in the larynx during phonation of voiced sounds is called the **fundamental frequency**. This is because it sets the periodic baseline for all higher-frequency harmonics contributed by the pharyngeal and oral resonance cavities above. The fundamental frequency also contributes more than any other single factor to the perception of pitch (the semi-musical rising and falling of voice tones) in speech. Vowels are an example of voiced sounds. In Figure 2.4, voiced sounds are represented by the pulse train generator, with the pitch (i.e., the fundamental frequency of the waveform) being an adjustable parameter.

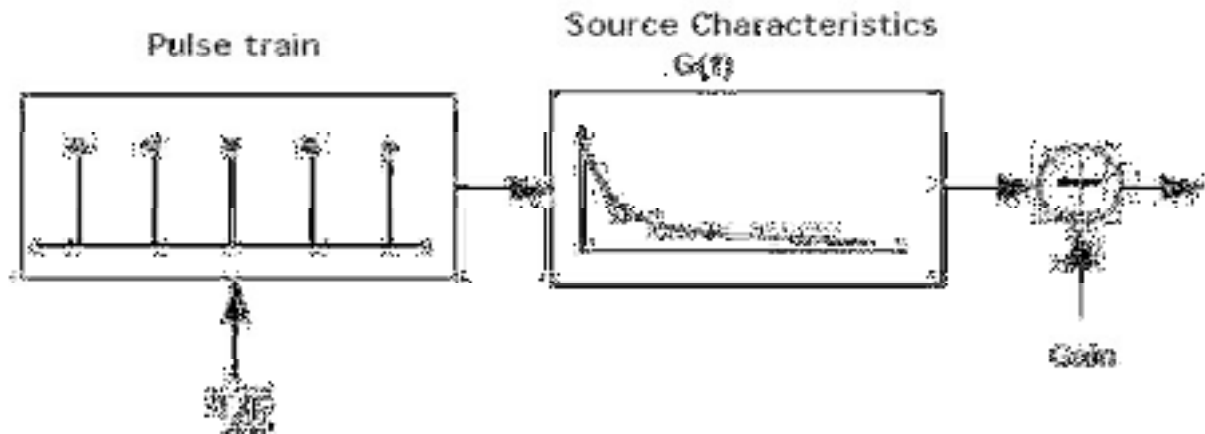


Figure 2.1 Block diagram of a basic voicing source model [29]

The model is composed of an impulse train generator that produces pulses at the rate of one per fundamental period. This impulse excitation simulates the generation of acoustic energy at the instant of the opening of the vocal cords. This signal then drives a linear filter whose frequency response $G(f)$ (Glottal function) approximates the glottal wave form. The function $G(f)$ must be chosen so that it approximates accurately the spectrum of the source. Finally, a gain control device allows the adjustment of the voicing amplitude.

In contrast, voiceless (fricative) sounds originate as random noises but not from vibration of the vocal cords. The vocal cords will be in no vibrating mode and are held open. This occurs when the air flow is nearly blocked by the tongue, lips, and/or teeth, resulting in air turbulence near the constriction. This phenomenon is due to a pressure drop across a constriction formed in the vocal tract, where the flow of air becomes turbulent [28] [29] In Figure 2.4 the fricatives are indicated by the voice generator.

Since the glottal wave is periodic, consisting of fundamental frequency (F_0) and a number of harmonics (integral multiples of F_0), it can be analyzed as a sum of sine waves. The resonances of the vocal tract (above the glottis) are excited by the glottal energy. Usually, the vocal tract is assumed as a straight tube of uniform cross-sectional area, closed at the glottal end, open at the lips. Depending on the shape of the acoustic tube (mainly influenced by tongue position), a sound wave traveling through it will be reflected in a certain way so that interferences will generate resonances at certain frequencies. These resonances are called formants and the frequency of these peaks of energy is called formant frequency. By changing the relative position of the tongue and lips, the formant frequencies can be changed in both frequency and amplitude. [28]

2.5 Basics of Text-to-Speech (TTS) Synthesis

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. [30] Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific use age domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output [31]. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1990s.



Figure 2.2 Stephen Hawking [32]

Stephen Hawking is one of the most famous people using speech synthesis to communicate and he (8 January 1942 – 14 March 2018) was an English theoretical physicist, cosmologist, and author who, at the time of his death, was director of research at the Centre for Theoretical Cosmology at the University of Cambridge. [32] Between 1979 and 2009, he was the Lucasian Professor of Mathematics at the University of Cambridge, widely viewed as one of the most prestigious academic posts in the world.

TTS is defined as the creation of speech by machines by way of the automatic phonetization of the sentences to complete [33]. Speech synthesis is a process of building the system that can generate human-like speech from any text input to mimic human speakers. [34] The ultimate objective of Text-to-Speech (TTS) synthesis systems is to create applications which listeners and users in general cannot easily determine whether the speech he or she is hearing comes from a human or a synthesizer. [34] This could possibly assert that, the ideal speech synthesizer should possess both high intelligibility and high naturalness of synthesized speech to achieve its crucial objective. Naturalness and intelligibility of speech are applied to the description of quality of speech synthesis system. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. The text-to-speech (TTS) synthesis procedure contains a number of steps and synthesized speech can be produced by different methods.

Natural Language Processing (NLP) and Digital Signal Processing (DSP) are generally two main phases of TTS systems in the process of converting written text into speech. [33] The former one is targeted to produce phonetic transcription of the text, together with the desired intonation and rhythm. This phase is also known as high-level synthesis [7]. The later one is transforms the symbolic information it receives from the former phase into speech [7]. This phase is also known as low-level synthesis. [7] Any TTS system contain the a below two phases as show in figure 2.6 [35]

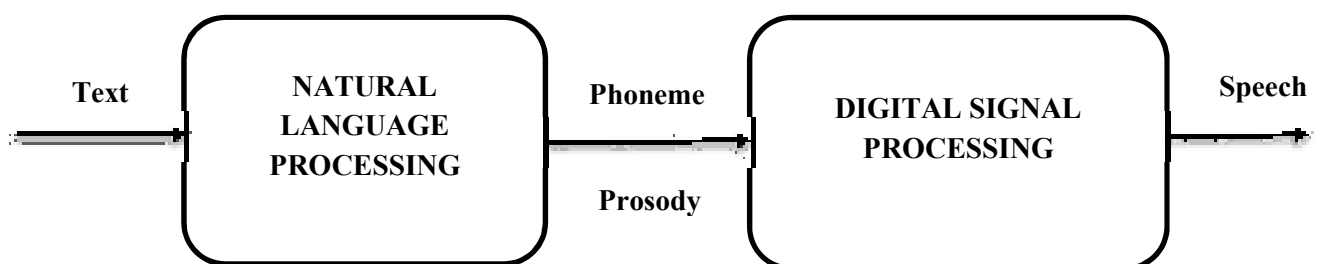


Figure 2.3 Basic Architecture of a TTS system [35]

2.6 Natural Language Processing (NLP)

One of the key steps of speech synthesis is the NLP (Natural language processing). The main activity in this phase is the text analysis and phonetics analysis. Also this part of the process is often the complex one of the two components. Numerals, abbreviations, and acronyms all need some preprocessing techniques to convert them into the corresponding full words. Correct prosody and pronunciation analysis from written text is also a difficult task because

written text does not contain explicit emotions that are expressed while speaking [2]. The task of converting the input text into a linguistic representation can be further partitioned into two components the transformation of text into phonetic units and the conversion of text into prosodic parameters. [9]

2.6.1 Text Analysis

Text analysis is all about transforming the input text into a 'speakable' form. At the minimum this contains the normalization of the text so that numbers and symbols become words abbreviations are replaced by their corresponding whole words or phrases and so on. This process typically employs a large set of rules that try to take some language-dependent and context-dependent factors into account. The most challenging task in the text analysis block is the linguistic analysis which means syntactic and semantic analysis and aims at understanding the content of the text. A computer cannot understand the text as humans do, but statistical methods are used to find the most probable meaning of the utterances. This is important because the pronunciation of a word may depend on its meaning and on the context (for example, the word record is pronounced in different ways depending on whether it is a verb or a noun). Lastly, the text analysis block is supposed to provide prosodic information to the subsequent stages and indicate the positions of pauses based on the punctuation marks, and distinguish interrogative clauses from statements so that the intonation can be adjusted accordingly.

2.6.2 Phonetic Analysis

Phonetic analysis converts the orthographical symbols into phonological ones using a phonetic alphabet. International Phonetic Association (IPA) contains not only phoneme symbols but also diacritical marks and other symbols related to pronunciation. Since the IPA symbols are rather complicated and there are several symbols that cannot be found in typewriters, other phonetic alphabets have also been developed. They are better compatible with computers and often based on ASCII characters. However, there is no generally accepted, common phonetic alphabet and therefore separate speech synthesizers often use their own special alphabets. [2] The degree of challenge in phonetic analysis is strongly language dependent.

2.6.3 Prosodic Analysis

The last stage of (Natural language processing) NLP is prosody generation also Prosodic features are said to be supra segmental. The term prosody is refers to elements of speech that are not individual phonetic segments vowels and consonants but are properties of syllables and larger units of speech, including linguistic functions such as intonation, tone, stress, and rhythm. [1] Prosody may reflect various features of the speaker or the utterance the emotional state of the speaker the form of the utterance statement, question, or command the presence of irony or sarcasm emphasis contrast and focus. It may otherwise reflect other elements of language that may not be encoded by grammar or by choice of vocabulary. [2]

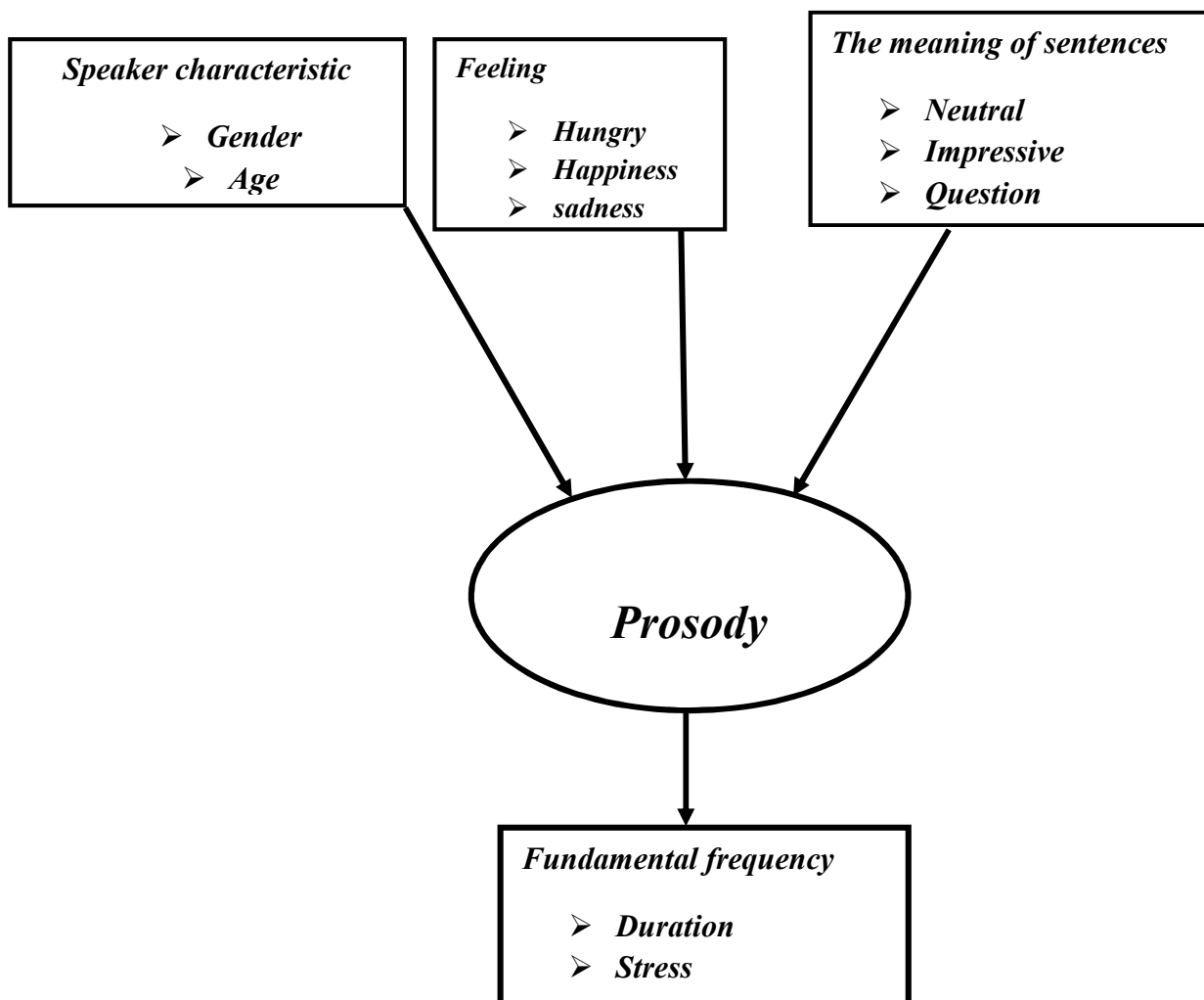


Figure 2.4 Factors contributing to prosodic feature [2]

2.7 Digital Signal Processing (DSP)

One of the second key steps of speech synthesis is the Digital signal processing (DSP). It is the use of digital processing, to perform a wide variety of signal processing operations. Generally it transforms the symbolic information (such as, phonetic transcription and prosodic information) it receives from the NLP phase into speech. [10] [7] There are many methods to produce speech after text and prosodic analysis. Usually the methods are classified in to three groups: articulatory, format and concatenative methods [14]. Articulatory synthesis attempts to model the human speech production system directly. Formant synthesis, which models the pole frequencies of speech signal transfers function of vocal tract into based on source-filter-model. Concatenative synthesis uses different length of prerecorded samples derived from natural speech. [10]

2.8 Speech synthesis techniques

The synthesis of a speech can be seen broadly from two major perspectives Natural point of view and the synthetic. The Natural speech synthesis is what humans do while communicating and the speech that called Artificial can be synthesized in a number of ways (Articulatory, formant or concatenative).

2.81 Articulatory Synthesis

Articulatory synthesis refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there. The first articulatory synthesizer regularly used for laboratory experiments was developed at Haskins Laboratories in the mid-1970s by Philip Rubin, Tom Baer, and Paul Mermelstein. [4] Articulatory synthesis is by far the most complicated in regard to the model structure and computational burden. The idea in articulatory synthesis is to model the human speech production mechanisms as perfectly as possible. Experiments with articulatory synthesis systems have not been as successful as with other synthesis systems but in theory it has the best potential for high-quality synthetic speech. [7]

The articulatory model also enables more accurate transient sounds than other synthesis techniques. Articulatory synthesis systems contain physical models of both the human vocal tract and the physiology of the vocal cords. It is common to use a set of area functions to model the variation of the cross-sectional area of the vocal tract between the larynx and the lips. The principle is thus similar to the one that has been seen within the acoustic tube

model. The articulatory model involves a large number of control parameters that are used for the very detailed adjustment of the position of lips and tongue, the lung pressure, the tension of vocal cords, and so on. The data that is used as the basis of the modeling is usually obtained through the X-ray analysis of natural speech. [7] As expected, such analysis is also very troublesome.

2.8.2 Formant Synthesis

Formant synthesis is the other method used to produce synthesized speech. In this synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model (physical modeling synthesis). [36] Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis.

Many systems based on formant synthesis technology generate artificial, robotic-sounding speech that would never be mistaken for human speech. However, maximum naturalness is not always the goal of a speech synthesis system, and formant synthesis systems have advantages over concatenative systems. Formant-synthesized speech can be reliably intelligible, even at very high speeds, avoiding the acoustic glitches that commonly plague concatenative systems. High-speed synthesized speech is used by the visually impaired to quickly navigate computers using a screen reader. Formant synthesizers are usually smaller programs than concatenative systems because they do not have a database of speech samples. They can therefore be used in embedded systems, where memory and microprocessor power are especially limited. Because formant-based systems have complete control of all aspects of the output speech, a wide variety of prosodies and intonations can be output, conveying not just questions and statements, but a variety of emotions and tones of voice. [37]

2.8.3 Concatenative Synthesis

Concatenative synthesis is based on stringing together or concatenating recorded voice clips. Concatenative synthesis typically results in the most realistic-sounding synthetic speech. However, there might occasionally be audible errors in the output due to variances between the nature of the automated procedures for segmenting the waveforms and the fluctuations in speech that occur naturally. Concatenative synthesis can be broken down into three primary subtypes (32). This is the so-called cut-and-paste synthesis, in which brief speech chunks are

chosen from a database of previously recorded audio and linked one after another to create the desired utterances. The use of real speech as the foundation for synthetic speech has the potential to produce speech of extremely high quality, but in practice there are significant restrictions, primarily because such a system needs a lot of memory. The synthetic speech will have fewer problematic concatenation sites as the chosen units get longer, but memory needs will also rise. Concatenative synthesis is also constrained by the output speech's substantial reliance on the selected database. For instance, it can be difficult to regulate a speaker's mood or personality. Concatenative synthesis, despite its somewhat featureless character, is ideally suited for a few specific, niche applications. Phonemes and Diphones are the most popular options since they are brief enough to achieve adequate flexibility and to maintain appropriate memory requirements. It is hard or impracticable to use longer units, such syllables or words, for a variety of reasons. Because a diphone contains the transition from one phoneme to another, the latter half of the first phoneme, and the former half of the latter phoneme, it offers some pretty good opportunities to account for co-articulation.

2.8.4 HMM Based Synthesis

HMM-based synthesis, often known as statistical parametric synthesis, is a synthesis technique based on hidden Markov models. In this approach, HMMs are used to simultaneously represent the vocal tract's frequency spectrum, the voice source's fundamental frequency, and the length of speech (prosody). HMMs themselves produce speech waveforms based on the maximum probability standard.

2.9 Speech Synthesis System

2.9.1 Festival Speech Synthesis System

The Centre for Speech Technology Research (CSTR) at the University of Edinburgh is where Alan W. Black, Paul Taylor, and Richard Daley first created the Festival Speech Synthesis System, a general multilingual speech synthesis system. It provides a comprehensive text-to-speech system with a number of APIs, as well as a setting for the study and development of speech synthesis methods. It is written in C++ and has a command interpreter that is similar to Scheme for general customization and extension.

2.9.2 Festvox Speech Synthesis System

Festvox is a suite of tools by Alan W. Black and Kevin Lenzo for building synthetic voices for Festival. The project aims to make the building of new synthetic voices more systematic and better documented making it possible for anyone to build a new voice. It is distributed under a free software license similar to the MIT License.

2.10 Related Works

Some great research has been done on speech synthesis for the local languages, taking the benefits of speech synthesis systems into consideration. Even though none of the studies used Hidden Markov Model (HMM) as their speech synthesis technique, they all set a benchmark for understanding fundamental ideas and introducing various speech methods for the preparation. In the following sections, different works related to speech synthesizers for local languages and speech synthesis using HMM for any language will be discussed. Both issues are related to this research either by the language behavior or the synthesis technique employed.

2.10.1 Text to Speech Researches on International Languages

English language speech synthesis also use the HMM method. [15] The authors have employed festival for text analysis and feature extraction, including contextual aspects, in their work. 524 sentences were used to train the model, and the speech data was captured at a rate of 16 kHz. Five left-to-rights HMM states make up their model. During speech synthesis, the contextual elements have been taken into consideration.

Though they did not put quantitative analysis, they have concluded that they have generated a natural sounding synthesized speech unlike any other rule based speech synthesizers, like formant based approach.

German speech synthesis is done using the HMM technique. The research in [16] explores how to apply HMM approaches to German expressive speech, focusing on how well HMM-based synthesis performs in the German language. The researchers evaluated two earlier studies that employed the same technique HMM-based synthesis to German—but they arrive to the conclusion that these studies only used a small number of training sets, leading to a small intelligibility. They attempted to integrate all of the German diphones in their effort to improve the intelligibility by using a speech database intended for the development of unit-selection algorithms that contains more than 3 hours of speech for each of the four speakers.

They employed a German speech synthesis corpus for training purposes. The 1683 sentences in this corpus were chosen from 2 male and 2 female speakers in order to provide the best coverage of the German diphone. They used professional recording equipment, a sound-proof environment with little reverb, a sampling rate of 48 KHz, and a resolution of 16 bits for the recordings.

Using a hidden Markov model (HMM) speech synthesis technique Ntsako et al [17] developed a highly intelligible and acceptable natural sounding speech synthesis system for the Xitsonga language of the Republic of South Africa. The evaluation process used mean opinion scores that ranged from one (worst) to five (best). The evaluation's findings show that, on the whole, the system was deemed to be satisfactory. Only 25% of the individuals from across the twelve languages thought the system was generally great, 37.5 percent thought it was good, and the remaining 37.5 percent thought it was adequate. This gave the system a 92.3% approval rating. On the other hand, the voice conversion mechanisms and algorithms employed throughout the speech generations were not described by the authors.

Concerning their final output, they have presented results about the adaptation of the synthesis system to a very limited set of expressive football comments and they have explained that the result they got is better than any other technique to have both flexible and natural sounding speech.

2.10.2 Text to speech Researches on Local Languages

Accordingly some remarkable works have been done on speech synthesis for local languages. As to the knowledge of this researcher none of the works done to the local languages used statistical parametric speech synthesis method using Festival as their toolkit. This section presents research works that uses different speech synthesis methods for local languages.

A research report of Tewodros Abebe [8] entitled Text- to-speech synthesizer for wolaytta language using speech synthesis architecture of Festival. This TTS (text-to-speech) is based on diphone concatenative synthesis, applying Residual LPC technique. Diphones are used as the basic concatenation units to synthesize. The test results indicate that the majority of the words are recognizable. The overall performance of the system is found to be 78% accurate.

Alula Tafere developed a generalized approach to Amharic text-to-speech (TTS) synthesis system. [18] This study has described generalized Amharic Text-To-Speech (TTS) synthesis, which attempt to handle both Amharic SWs and NSWs. The system is developed using

speech synthesis framework of Festival, based on diphone unit concatenative synthesis by applying RELP coding technique. The performance of the system shows that on the average 73.35% words both SWs and NSWs correctly pronounced.

The same methodology as that of Laine was addressed in a research paper by Henok [38], but Henok employed time-domain Pitch Synchronous Overlap and ADD (PSOLA) technique rather than linear predictive methods (LPM) to produce the synthetic speech. Additionally, he has taken into account prosodic effects like rage, joy, and emotions, which Laine did not. Additionally, Habtamu [39] uses diphones as a speech unit to implement the concatenative speech synthesis technique.

For the Tigrinya language, Agazi Kiflu [20] created a unit selection-based text to speech synthesizer. According to the findings the author analyzed, 2.7% thought the voice sounded weird, 58.3% thought the voice was good, and 38.8% thought it was very good. The performance as a whole is higher than 97.1%, which is a respectable rate. However, a unit selection technique needs a sizable database, frequently in the gigabyte (GB) range, depends for a single speaker, and a constrained domain vocabulary.

Lemlem H and Million M [40] developed concatenative based text to speech synthesizer for Tigrinya language. They developed the prototype using festival speech synthesis framework that resulted 89.76% using 15 test case utterances (user based evaluation). Their synthesizer was evaluated via user based evaluation which is prone to high error rate. All these mentioned synthesis types are concatenative speech synthesis systems and their advantages of these techniques is high quality speech but it is disadvantageous due to enormous costs and time for constructing corpora and is not straightforward to synthesize diverse speakers such as emotions, styles, database dependent and only works for limited domain.

Nadew's speech synthesis for Amharic [41], His work synthesizes Amharic vowels using a rule-based method called formant-based speech synthesis, in contrast to earlier research that employed data-driven techniques.

Samson Tadesse [19] developed a concatenative based text to speech synthesizer for Afaan Oromoo language. During the process, a limited rule based diphone database entries were constructed. The author showed that 75% and 54% of words in the data set are correctly pronounced as to the diphone and triphone speech units respectively, i.e., the concatenation of large units degraded the performance of the system. Based on mean opinion score

measure, the author achieved the intelligibility results, 3.03 and 2.2 rate for the diphone and triphones respectively and the naturalness of the system was 2.65 and 2.02 for each speech units respectively. However, due to the use of a rule based approach, the overall performance of the system is became poor. Although, the work of the author has a lot of limitations such as it require huge memory, database dependent and also needs a larger data base to include all possible utterance in the language.

A research work by Tesfay Yihdego [42] developed a prototype for Tigrigna Language using TTS System that is Diphone based concatenative speech synthesis. Diphones are used as the basic concatenation units to synthesize sample Tigrigna texts and also Time Domain Pitch Synchronous Overlap and Add (TDPSOLA) used as a technique to generate the synthetic speech.

Bereket Kasseye [43] developed a text to speech synthesizer for Amharic language using hidden Markov model technique. The results from the mean opinion scale (MOS) were found to be 4.12 and 3.6 for intelligibility and naturalness respectively. However, the study did not consider the factor of intonation in developing the system.

We reviewed a variety of speech synthesizers that were created with a focus on various methodologies in this chapter. Rule-based and machine learning-based techniques are introduced and debated. Learning using hidden Markov models has demonstrated some intriguing results. Learning based on hidden Markov models provides the capacity to capture Speech synthesis issues. For instance, Ntsako's efforts [17]support the effectiveness of HMMs

Table 2.1 some related work in the form of table

Sr. No	year	Author(s)	Focus of the paper	Research Gaps
1.	2002	K. Tokuda, Heiga Zen, Alan W. Black	An HMM based speech synthesis system applied	they did not put quantitative analysis
2.	2004	Agazi Kiflu	Unit Selection Based Text-to-Speech Synthesizer for Tigrinya Language	It needs a sizable database, frequently in the gigabyte (GB) range,
3.	2009	Tewodros Abebe	Text-to-Speech Synthesizer for Wolaytta Language	This TTS (text-to-speech) is based on diphone concatenative synthesis, applying Residual LPC technique.
4.	2010	Tafere, Alula	A Generalized approach to Amharic text to speech	the author recommendation better to shift from rule based approach to statistical parametric based approach
5.	2011	S. Tadesse	Concatenative Text-To-Speech System for Afaan Oromo Language	it require huge memory, database dependent and also needs a larger data base
6.	2015	B. Ntsako	Text-To-Speech Synthesis System for Xitsonga using Hidden Markov	The voice conversion mechanisms and algorithms employed throughout the speech generations were not described

2.11 Kambaata language and Phonology

2.11.1The Kambaata People and the language

Kambaata is in the Southern Nations, Nationalities, and Peoples Region (SNNPR). They live in the highland areas around the Abarrichcho massif, about 300 kilometers southwest of the Ethiopian capital, Addis Ababa, between the Omo River to the west and the Billate River to the east. The Kambaata-Tembaaro Zone comprises three major ethnic groups: Kambaata, Tembaaro, and Donga. The capital of the zone is Duuraame, 300 kilometers from Addis

Ababa. According to the 2007 Ethiopian housing and population census, the number of Kambaata speakers is 685, 167, among them 337, 852 males and 345, 315 females. [14]

The zone contains three city administrations and seven wärādas (districts). The three city administrations are Duuraame city administration, Shiinshichcho city administration, and Hadaro city administration. The seven wärādas are Qadiida Gaameela wärāda, Qaacca Biira wärāda, Angacca wärāda, Tembaarowärāda, Haddaro and TuntoZuriya wärāda, Daambooyya wärāda, and Dooyyeeganna wärāda. [14]



Figure 2.5 Kembata tembaro administrative map [14]

The Cushitic language family is split into four sections, according to Grimes (2000) and the web edition of Ethnologies: Central Cushitic, East Cushitic, North Cushitic, and South Cushitic. The Highland East Cushitic language group is one of the branches of the East Cushitic language family. [44]

Despite the fact that the sources mentioned above list seven languages as Highland East Cushitic languages (Alaaba, Burji, Gedeo, Hadiyya, Kambaata, Libido, and Sidaama), only five of them are usually labeled as such in the literature (e.g. Fleming and Bender 1976, Hudson 1976), where Libido is treated as a dialect of Kambaat and Alaaba as a dialect of Hadiyya Hudson (1976:236-246) summarizes the discussions and issues surrounding Highland East Cushitic language categorization, particularly the position of Burji.

The position of Kambaata within the branch of Cushitic is shown in the following family tree:

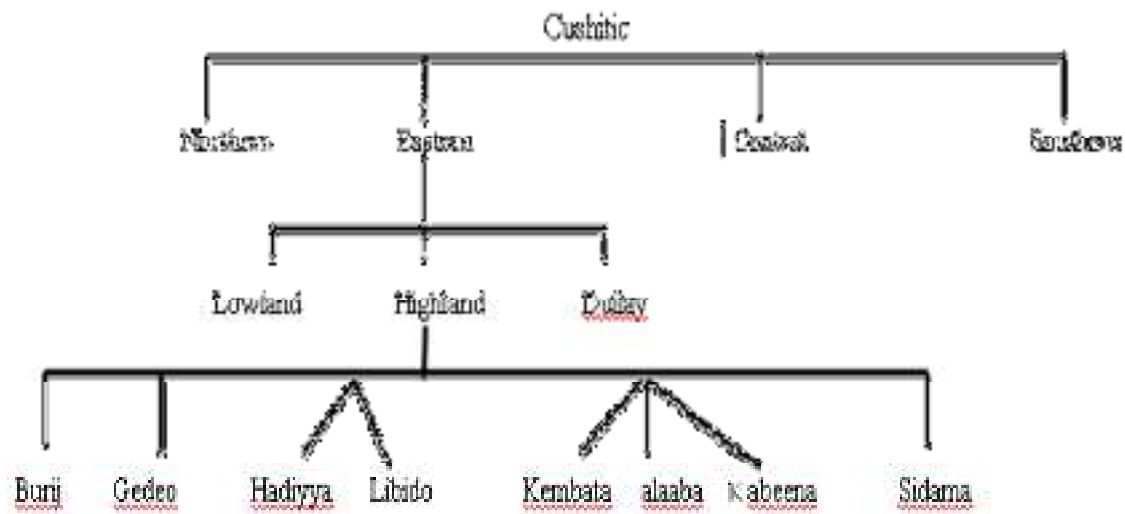


Figure 2.6 The branch of Cushitic family [14]

In addition to being the name of a single language, Kambaata is also the name of a minor HEC branch that includes the languages Kambaata (in the strict sense), Tembaaro, Alaaba, and Qabeena. According to [14] Alaaba and Kambaata are unlike in some lexical areas and grammatical constructions. On the other hand, she emphasizes that although there are slight distinctions between Kambaata and Tembaaro, they are both comprehensible. The closest relative of Qabeena is Alaaba Crass who was cited by Treis (2008:4). Scholars have differing opinions about how to classify (HEC). It has been a difficult matter to include Burji in the (HEC) branch (Hudson 1976: 241-244). Because of this, most academics classify Burji as a different branch of HEC. E.C. Kambaata has been the major language of instruction in primary schools since 1993. It has been taught as a subject in high schools since 2007 E.C. the Latin-based official orthography. [14]

2.11.2 Phonology

This article identifies and describes Kambaata speaking sounds. Furthermore, the system's shared phonological processes, syllable structure, and co-occurrence limitations of segments in the language are discussed. The IPA conventions as amended in 1993 are primarily used in the transcription.

Phonetic symbols that resemble letters in alphabetic languages like English are used to represent speech sounds called phones. As an illustration, a phone usually corresponding to

the letter f is represented by the symbol f, while a phone typically corresponding to the letter g is represented by the symbol g. Some phones may have two or more characters to represent them in some languages. For the Kambaata language, the three most prevalent single phones are ch, ph, and sh [16]. Phonology, in its simplest form, is the study of how sounds are arranged and employed in natural languages using sounds, vowels, and consonants. [14]

Aa	Bb	Cc	Dd	Ee	Ff	Gg	Hh	Ii	Jj	Kk
/a/	/b/	/tʃ̣/	/d/	/e/	/f/	/g/	/h/	/i/	/dʒ/	/k/
Ll	Mm	Nn	Oo	Qq	Rr	Ss	Tt	Uu	Ww	Xx
/l/	/m/	/n/	/o/	/kʰ/	/r/	/s/	/t/	/u/	/w/	/tʰ/
Yy	Zz	CH	ch	PH	ph	SH	sh	’		
/j/	/z/	/tʃ/	/Pʰ/	/ʃ/	/ʔ/					

Consonant phonemes

In the following, each consonant phoneme will be treated separately. For each phoneme, firstly, examples of occurrence in word-initial and intervocalic position as well as in a cluster of identical and different consonants will be given. Secondly, minimal or near minimal pairs will be provided to illustrate the phonemic contrast between simplex and Geminate consonants. Note son distribution restrictions will be given. [45]

Stops

Three types of stops are distinguished in Kambaata: voiced, voiceless, and glottalic(ejective) stops.⁶ They are produced at five places of articulation: labial, alveolar, palato-alveolar, velar, and glottal. Voiceless stops are slightly aspirated, but aspiration is not a phonemic feature of Kambaata. In the palato-alveolar region the oral closure is not opened by a burst; instead, all palato-alveolar stops are phonetically affricates. Geminate stops are generally characterized by an extended closure time. Through this, more pressure is built up and the burst of the closure becomes more intense than the one of a simplex stop. Like many Cushitic and Ethio-Semitic languages, Kambaata has no voiceless bilabial plosive /p/ in native words. ⁷The phoneme /p/ is, however, found in some loanwords that Kambaata borrowed from European languages via Amharic; see /po(o)stímini/ ‘post office’, /ampulá/ ‘light bulb’, /pappaayyá/ ‘papaya’ or /paastá/ ‘pasta’. [45]

/b/ is a voiced bilabial plosive.

- /buurú/ ‘butter’, /habará/ ‘enset leaf’, /bubbíta/ ‘whirlwind’, /dimbú/ ‘tobe(come) drunk’
- /b/ — /bb/ /zabú/ ‘to hold back, hold up’ — /zabbú/ ‘medicin

Intervocalic lenition is very prominent for the phoneme /b/. In the production of intervocalic /b/ very little if any pressure is built up after the oral closure. In an almost approximant-like manner, the lips close only softly before they open again. The difference between intervocalic /b/ and /bb/ is therefore not solely a difference between a short and long realization of the consonant, but also a difference in the mode of articulation.

/p’/ is a bilabial ejective stop with a defective distribution. In native words it does not occur word-initially

- /t’up’á/ ‘pestle’, /buup’p’á/ ‘central leaf sheath of enset plant’, /emp’eep’p’ú/ ‘to stretch to reach’
- /p’/ — /p’p’/
/k’op’á/ ‘milk of the first day after the birth of a calf’ /k’op’p’aná/ ‘lie’/óp’i/ ‘Climb!’, (2SG.IMP) /óp’p’i/ ‘(I) having climbed’ (1SG.PCO)

In the Amharic loanword for the 13th month of the Ethiopian calendar, /p’aaguumeéta/, the initial /p’/ is often replaced by /k’/ – as is also the case in many Amharic dialects.

/t/ is a voiceless alveolar plosive.

- /tassáa/ ‘hope’, /huggaati/ ‘whey’, /faattáta/ ‘enset leaf without the midrib’, /galtíta/ ‘residence’
- /t/ — /tt/
/alitá/ ‘vermin’ — /alittáta/ ‘vermin’ (PL1)

/d/ is a voiced alveolar plosive.

- /duuná/ ‘mountain’, /odoorrá/ ‘acacia tree’, /addá/ ‘truth’, /buundalá/ ‘bumble-bee’
- /d/ — /dd/
/madá/ ‘big dish, bowl’ /maddá/ ‘shroud, burial wrapping’

/t'/ is an alveolar ejective stop. It regularly replaces the Amharic ejective fricative/s'/ in loanwords (42).

- /t'eená/ 'rain', /fiit'á/ 'honey for ritual purpose', /k'iit't'ú/ 'to pity, have mercy', /k'int'á/ 'one quarter of an enset bread'
- /t'/ — /t't'/
 /t'aat'ú/ 'to wrap' — /t'aat't'ú/ 'to be ready, to be prepared'
 /k'at'á/ 'amount, extent' — /k'at't'á/ name of a mountain in between Duuraameand Daambooyya

Amharic Kambaata

'book' mäs'haf (መልህፍ) mat'aafá

/tʃ/ is a voiceless palato-alveolar affricate with a slightly defective distribution.

There are only a few lexemes starting with the phoneme/tʃ/; most of them are ideophones.

The intervocalic occurrence of simplex tʃ is even rarer, which might have led other authors to assume earlier that /tʃ/ can only appear in geminate form in intervocalic position. [45]

- /mik'áta/ 'bones', /mik'it tʃ tʃ ú/ 'bone'

/dʒ/ is a voiced palato-alveolar affricate.

- /wodʒdʒú/ 'white', /in dʒiidʒdʒíta/ 'tears'

/k/ is a voiceless velar plosive. Non-geminate intervocalic /k/ is not frequent in native Kambaata words; but it is often attested in loan words, e.g. /aakiimá/ 'doctor' and /abokaatúta/ 'avocado'.

- kohá/ 'guest', /atakaanú/ 'type of enset dish', /bokkú/ 'relatives, major lineage', /haankurú/ 'tosteam'

/g/ is a voiced velar plosive.

- /geinú/ 'yoghurt', /ragáú/ 'to inherit', /iggá/ 'bold, daring, courageous', /lungá/ 'weak'
- /g/ — /gg/
 /wogá/ 'custom, culture' — /woggáa/ 'year'
 /lugumú/ 'origin, descent' — /luggumú/ 'type of vessel'

/k'/ is a voiceless velar ejective.

- /k'eessá/ 'cheese', /dak'ayyá/ 'pasture, large meadow', /wok'k'áa/ 'road', /t'enk'úta/ 'mug'
- /k'/ — /k'k'/
/tak'íta/ 'cause, means' — /hak'k'íta/ 'wood'
/k'ook'áta/ 'blind' (F) — /k'ook'k'áta/ 'blind'

/ʔ/ is a voiceless glottal stop. The distribution of the glottal stop is the same as that of other consonant phonemes, i.e. it occurs in word-initial position, intervocalically, and in clusters with a sonorant or a glide. The glottal stop differs from other obstruents with respect to the position that it occupies in clusters: it must be the first (rather than the second) component.

- /doʔná/ 'nest', /hooʔmíta/ 'passion fruit'

Fricatives

Kambaata distinguishes voiced and voiceless fricatives. Voiceless fricatives are phonemic at four places of articulation: labial, alveolar, palato-alveolar, and glottal. In the voiced domain, the number of phonemes is reduced to two, or actually to one safely established phoneme and one disputable phoneme, if the highly defective distribution of the voiced palato-alveolar fricative

/ is taken into consideration. Unlike the reconstructed PEC, but like other HEC languages, Kambaata does not have pharyngeal fricatives. [45]

/f/ is a voiceless labio-dental fricative.

- /fokkúta/ 'uncultured behavior', /k'ok'k'ofaadá/ 'wood-pecker', /t'effúta/ 'smallenset bread', /hamfarrúta/ 'much hair on cheeks, neck, chest (of men)'
- /f/ — /ff/
/hofáta/ 'deep hole' — /hoffáta/ 'Saturday'
/afóo/ 'mouth' — /affóo/ 'who takes'

/s/ is a voiceless alveolar fricative.

- /sulumúta/ 'heifer', /gisanáta/ 'sleep', /kambaatissáta/ 'Kambaata language', /tunsúta/ 'darkness'
- /s/ — /ss/

/dasú/ ‘to be late’ — /dassú/ ‘to chop (an enset corm)’

/z/ is a voiced alveolar fricative.

- /z/ — /zz/
/azúta/ ‘milk’ — /zazzalúta/ ‘trade’
- /z/ — /d/
/zaná/ ‘living fence’ — /daná/ ‘beauty’, /zirrú/ ‘to destroy’ — /dirrú/ ‘to descend’

/ʒ/ is a voiced palato-alveolar fricative with a highly defective distribution. It is not realized as a simplex phoneme, but only occurs in its geminate form intervocalically.

- /sóoʒʒeeu/ ‘it dawned’
The discussion about its phoneme status is postponed to the section on morph phonology, as /ʒ/ is the result of a morph phonological process in one regional lect of Kambaata.

/ʃ/ is a palato-alveolar fricative.

- /ʃoolú/ ‘four’, /kiʃ á/ ‘hip’, /maʃ ʃáata/ ‘big knife (for enset food)’, /wonʃ úta/ ‘filter’

/h/ is a voiceless glottal fricative, which occurs word-initially and in non-geminate form intervocalically.

- /haamúta/ ‘chest (of humans)’, /bahá/ ‘artificial calf’

Nasals

Kambaata has two safely established phonemic nasals, /m/ and /n/. The phoneme status of the palato-alveolar /ɲ/ is very doubtful. Geminate nasals simply have a longer oral closure.

/m/ is a bilabial nasal.

- /maassáta/ ‘blessing’, /muummí/ ‘hair (of humans)’, /dimbaabíta/ ‘umbrella’
- /t’umá/ ‘good, nice’ — /t’ummá/ ‘peace’

/n/ is an alveolar nasal.

- /nubaabú/ ‘old (wo)men’, /borkaanú/ ‘neck-rest’
- n/ — /m/

/dúnu/ ‘pouring away

Liquids

The term “liquid” refers to the class of laterals (“l-sounds”) and vibrant s / taps (“r sounds”) (Hall 2000:84). Liquids in Kambaata make a phonemic distinction between plain and glottalized. This phonemic distinction has eluded the attention of linguists who have worked on Kambaata before. I am not aware of descriptions asserting the existence of such consonants in other Cushitic languages. [45]

/r/ is an alveolar trill. A geminate intervocalic /rr/ is realized as a long trill, i.e. many vibrations of the tip of the tongue. A simplex intervocalic /r/ is subject to lenition.

- /reek’k’éeta/ ‘obsidian rock’, /eeráa/ ‘enset plant (3rd stage)’, /boorrasú/ ‘to feel bored’, /murtíta/ ‘decision’.
- /r/ — /rr/
/birú/ ‘to praise’ — /birrú/ ‘to stop (of rain)’

/rʔ/ is a glottalized / laryngealized alveolar tap with a defective distribution. It only occurs between vowels or as the first element of a cluster. All the words known so far as containing this peculiar phoneme are given in (93). As most /rʔ/-lexemes have a very specialized meaning the translation of the following words should only be considered an approximation.

- zur’á/ n. ‘ear of grain, ear of coffee’
- /ziir’á/ ‘vermin in the leaf sheaths of enset plants’

/l/ is a lateral approximant during whose production the tip of the tongue touches the alveolar ridge. Geminate /ll/ simply has a longer oral closure than a simple one.

- /lelléeta/ ‘pumpkin’, /laláta/ ‘soot’, /gaallá/ ‘thigh’, /kaltá/ ‘hatchet’, /ukkulá/ ‘hoof’

/lʔ/ is a glottalized / laryngealized lateral approximant with distribution as defective as /rʔ/. It is never attested word-initially. The few words with a glottalized /lʔ/ that have been found so far are given in

Lexemes with /lʔ/

/gal’á/ n. ‘shard’

/d al'á/ *adj.* 'sloppy (e.g. of bean or pea pods that are not fresh any more)'

/d al'íta/ *n.* 'crème, lotion'

Glides

/w/ is a labio-velar approximant, i.e. an approximant with two places of articulation. In its production the lips approach each other and the back of the tongue is raised towards the soft palate

/w/ — /ww/

/hawaandaá/ 'time of high heat in the afternoon when worker bees return to their hive' —
/hawwá/ 'problem'

After word-initial /w/, the quality of short non-high vowels is difficult to determine. The transcription of wV-words in the literature often differs from author to author; alternating transcriptions can even be found in the work of one and the same author. See for example three examples from the HEC dictionary [45]

- /wa'a/ ~ /wi'a/ 'water'
- /wuk'k'a/ ~ /wok'k'a/ 'road'
- /wad-/ ~ /wood-/ 'roar'

/y/ is a palato-alveolar approximant which according to IPA conventions would have to be transcribed as [j]

/yaaú/ 'to hold a meeting', /ayé/ 'who?', /geegeeyyá/ 'gift of the bride's parents for the groom', /baayaatú/ 'to mention'

/woyiná/ < wäyn < ENGL vine

/guwantá/ < gwant < FR gant, ITAL guanto 'glove'

/y/ — /w/

/yaarú/ 'to cry' — /waalú/ 'to come', /yaallá/ 'serious disease' — /waallé/ 'Come!'

Generally, Kambaata has 25 consonant phonemes which are shown in Table

Table 2.2 Consonant phonemes

Point to articulation		Labial	Labio-dental al	Alveolar	Palatal	Velar	Glottal
Manner of articulation							
Stops	Voiceless			T	ʃ	k	ʔ
	Voiced	b		D	dʒ	g	
	Glottslie	pʼ		tʼ	cʼ	kʼ	
Fricative	voiceless		F	S	ʃ		H
	voiced			Z	(ʒ)		
Nasal		M		N	(ɲ)		
Liquid	Lateral			L			
	Trill			R			
Glide		W			Y		

Vowels

The occurrence of vowels in word-initial position

Table 2.3 the occurrence of vowels in word-initial position

Vowels	Idiophones	Gloss
/i/	Ilil	Ululate
/e/	ekkʼu	Okay
/o/	oʼorin-oʼorin	stimulate a child who drinks koso
/u/	ukʼuk	give extra care for sb/sth
/a/	aʼakk	Sound made by spitting a chest cold

The occurrence of vowels in word-medial position

Table 2.4 the occurrence of vowels in word-medial position

Vowels	Idiophones	Gloss
/i/	Libb	think, remember or recollect
/e/	c'eeek'k	glint of momentary light
/o/	fo'ott	lift sth quickly
/u/	k'uc'uc	scrunch or crash a bone
/a/	ba'tt	be confused

The occurrence of vowels in word-final position

Table 2.5 the occurrence of vowels in word-final position

Vowels	Idiophones	Gloss
/i/	C'ii	sound made by birds
/e/	k'émbe	of a young girl grow well
/o/	murt'ú	of fruit to be overripe
/u/	dordzó	wobble or move side ways
/a/	Fandá	become fat

All the five vowels appear in ideophones in word-initial, medial and final positions. Long vowels: front vowels /e/ and /a/ as well as back vowels /u/ & /o/ are employed in medial position

in ideophones such as *c'éeek'k* 'glint or momentary light', *fáak'k* 'sweep roughly', *húup'p* 'to drink a thick liquid' and *fóott* 'lift something quickly' respectively.

Syllable structure and phonotactics

A syllable always begins with a consonant or a vowel, and may end either with a short or a long vowel or a single consonant (or geminated consonant). The nucleus of the syllable contains either short vowel or long vowel. A syllable can be closed or open.

A syllable is a phonological unit which consists of one or more sounds. It is typically made up of a nucleus (most often vowel) with optional initial and final consonants. It is often considered the phonological —building block ‘of words. A syllable is one of the phonotactic constraints. The onset of every Kambaata syllable is occupied by a consonant. The syllable onset should not be empty, but the coda can be empty or be filled by a single consonant.

Types of syllable

With regard to types of syllable in non-ideophonic words, ám ‘come’, manǰ-ú ‘man’, kambáata ‘Kambaata people’ and wengereelliǰǰ-ú ‘fox’ are monosyllabic, disyllabic, trisyllabic and multisyllabic words respectively. [45] Identified four types of syllable structures in regular/non-ideophonic words: CV, CVV, CVC, and CVVC. For instance, the monosyllabic non ideophones such as kú ‘this’, mii ‘why’, már ‘go’ and múud ‘to remove the layer of a covering’ exhibit CV, CVV, CVC and CVVC syllable structure respectively. Ideophones have four types of syllables: monosyllabic, disyllabic, trisyllabic and multisyllabic ideophones, but they have five types of syllable structure:

- 1) CV: gú ‘a word used in dealing with sth. taboo’, tú ‘to spit saliva’
- 2) CVV: jáa ‘sound made by light rain’, búu ‘sound made by house fly’
- 3) CVC: táf, ‘to be restless’, kús ‘to walk gracelessly’
- 4) CVC¹C¹: bátt ‘be confused’, ǰinn ‘to laugh derisively’
- 5) CVC¹C²: ménk ‘to be talkative’, sirp ‘become silent/quiet’

Monosyllabic ideophones

The overwhelming majority of monosyllabic ideophones are sound-symbolic; they imitate the natural sound. The monosyllabic ideophones with long vowel demonstrate iconic lengthening. The monosyllabic ideophones best illustrate imitative sound symbolism. However, there are monosyllabic ideophones such as dǰúu ‘become confused’ and dúu ‘become bewildered’ are not sound symbolic in origin. In addition, the ideophone gú ‘a word used when talking about something regarded as taboo’ is not onomatopoeic. In contrast, ideophone tú ‘to spit saliva’ is sound symbolic. Furthermore, non-ideophonic monosyllabic words are scanty in the language.

For example, *mée* ‘give me this!’ *ʔáa* ‘mum’, demonstratives: *ká* ‘this’, ‘and *tá* ‘this’ as well as the possessive adjectives: *íi* ‘my’, *kíi* ‘your’ and *níi* ‘our’ are non-ideophonic monosyllabic words

Disyllabic ideophones

Disyllabic ideophones have the syllable structure of: CVCV, CVC¹C¹VC, CVC¹C²V, CVCVC, CVCVC¹C¹ and CVC¹C²VC¹C¹. The following sections deal with the phonological properties of this aspect. For instance, below some disyllabic ideophones have closed syllable, whereas others have open syllable, Note that all of the ideophones below have the syllable structure of CVCV or CVC¹C¹VC

hebó become crazy

redʒé ‘get mollified’

kaʃǎ ‘argue over sth.’

faggáʔ ‘wash or clean completely

burráʔ ‘make porridge’

Trisyllabic ideophones

Tri syllabic ideophones contain the syllable structure of: CVCVCVC, CVCVC¹C²V, CVC¹C²VC¹C¹VC and CVCVCVC¹C¹. Tri syllabic ideophones in the below the syllable structure of CVCVCVC or CVCVC¹C²V; however, tri syllabic ideophones such as *hambúk’k’aʔ* ‘to drink large amount of something hurriedly and *tonkóllaʔ* ‘gulp down or drink something in large amount hurriedly and completely have the syllable structure of CVC¹C²VC¹C¹VC. [45]

kiríraʔ ‘to rotate’

gagárda ‘of equines move fast’

susúk’aʔ ‘to push someone or something’

muk'úk'a? 'to open something by twisting rotating'

kulúla? 'make sb. feel giddy or dizzy'

Multisyllabic ideophones

Multisyllabic ideophones are ideophones that their syllable structure constitute four and above four syllable. The prefix 'multi' means 'many or several'. Some scholars use the term quadric syllabic for ideophones with four syllables; however, others use the term polysyllabic instead of multisyllabic and quadric syllabic. Ideophones with syllable structure more than four are not attested in the language. The multisyllabic ideo phone kuukkúlukuu __roaster's crow is sound symbolic, especially onomatopoeic and it contains long vowel which is used to show iconic meaning. Note that all of the multisyllabic ideophones below CVCVCVC'C'VC orCVC'C²VCVC'C'VC phonological template. [45]

jabálakka? 'to hedge or avoid giving direct reply'

t'uk'úrumma? 'to immerse into water'

munt'úluk'k'a? 'to do sth completely'

Morphological process

Morphology is the identification, analysis and description of the structure of morphemes and other units of meaning in a language like words, affixes, and parts of speech and intonation/stress, implied context. Morph phonological processes are especially prominent in the verbal domain. As it is discussed by it is a process that involves Assimilation, Epenthesis and Metathesis. [16]

Epenthesis

Epenthesis is observed whenever a stem ending in a consonant cluster connects with a consonant-initial person marker. The insertion of the epenthetic vowel /i/ helps to avoid sequences of three consonants and to keep the two-consonant constraint. The formula for epenthesis is given below

Epenthesis:

I) CC + t CC-i-t

II) CC + n CC-i-n

Below, verbal stems with clusters of identical or non-identical consonants are joined with various /t/- or /n/-initial inflection suffixes.

/toll-/ ‘stretch out’ + /-t-áau/ 3F/PL.IPV

➤ /tollitáau/ ‘she / they will stretch out’

/fint-/ ‘list’ + /-n-óommi/ 1PL.PVO

➤ /fintinóommi/ ‘we listed’

Metathesis

Metathesis refers to a phonological process that transposes two adjacent consonant phonemes. Metathesis is activated by attaching sonorant-initial morphemes, such as the /n/-initial 1PL person marker, to a base ending in an obstruent consonant (except glottal stop). The sonorant and the obstruent exchange their position as captured and exemplified below. Metathesis avoids unlicensed *O.S clusters and brings about authorized S clusters. [45]

Metathesis I: $O + S \rightarrow S.O$ (if $O \neq ?$)

/biit’-/ ‘break’ + /-n-eemmi/ 1PL.PVE \rightarrow /b’iint’eemmi/ ‘we broke’

/kaas-/ ‘plant’ + /-n-unta/ 1PL.PURP.DS \rightarrow /káansunta/ ‘so that we plant’

Metathesis is accompanied by two other automatic processes: (i) nasal place assimilation and (ii) fortition of /h/. Recall that clusters of a nasal and an obstruent (which is not a glottal stop) must be homorganic. After the metathesis of alveolar /n/ and a non-alveolar obstruent, e.g. bilabial /b/, the nasal obligatorily assimilates in place to the consonant then following

/bub-/ ‘burn’ + /-n-aammi/ 1PL.IPV

\rightarrow */bunbáammi/ \rightarrow /bumbáammi/ ‘we will burn’

Assimilation

Assimilation is the process in which a consonant segment becomes more alike its neighboring sounds. The direction of assimilation is from right to left, i.e. we are dealing with progressive or perseverative assimilation [45]

Assimilation: O + t → O.O

Assimilation is not triggered by verbal stems ending in a sonorant, because stem-final sonorant and suffix initial /t/ form a licensed cluster; see /mar-/ ‘go’ + /-tooti/ → /mártooti/ ‘Don’t go!’. Verbal stems ending in all possible obstruents are joined with the /t/-initial negative imperative singular morpheme, /-tooti/, in the examples

‘Don’t V!’

b: /ʔeeb-/ ‘bring’ /ʔéebbbooti/

p’: /k’op’-/ ‘throw’ /k’óp’p’ooti/

t: /sut-/ ‘insert’ /súttooti/

k’: /nak’-/ ‘hit’ /nák’k’ooti/

f: /hank’af-/ ‘embrace’ /hank’áffooti/

s: /mazees-/ ‘wound’ /mazéessooti/

z: /gaaz-/ ‘wage war’ /gáazzooti/

Chapter Three

Methodology and Materials

3.1 The Text to Speech Algorithm

The statistical parametric technique is used in the TTS system created for the Kambaata Language. The most popular statistical method of speech synthesis system from this perspective is HMM. It is the method that the majority of speech researchers utilize. The most popular algorithm now used to synthesis speech utilizing Festival architecture is one of the available several techniques.

3.2 Overview of statistical parametric synthesis

The most straightforward way to define statistical parametric synthesis, despite the fact that it produces some of the highest quality synthetic speech, is to generate the average of a few sets of speech segments with comparable tones. The use of a pre-recorded voice corpus, from which a selection of parameters is extracted, is implied by parametric corpus-based synthesis. [46] Speech synthesis thus turns into a statistical study of a speech corpus.

A typical statistical parametric voice synthesis system gathers spectral and excitation parameters as well as parametric representations of speech from a speech database, and then models those representations using a variety of generative models, such as hidden Markov models (HMMs). It is a concept that was taken from automatic speech recognition, and speech synthesis can also use it extremely well and with a lot of versatility. In most cases, the model parameters are estimated using the maximum likelihood (ML) criterion as

$$\lambda = \arg \max \{p(O|W, \lambda)\}$$

Where λ is a set of model parameters, O is a set of training data, and W is a set of word sequences corresponding to O . We then generate speech parameters, O , for a given word sequence to be synthesized, w , from the set of estimated models to maximize their output probabilities as

$$\hat{O} = \arg \max \{p(o|w, \lambda)\}$$

From the parametric representations of speech, a speech waveform is finally produced. It is possible to employ any generative model, but HMMs have been the most frequently utilized.

The term "HMM-based speech synthesis" refers to statistical parametric speech synthesis using HMMs. [47]

The advantages of the statistical parametric synthesis refer to

- ❖ Smooth and stable Averaging speech units
- ❖ Small footprint it Store statistics rather than waveforms
- ❖ Language independent focus Only on contexts and questions
- ❖ Easy to change style & emotions like Adaptation, interpolation

3.3 The Hidden Markov Model

HMM-based Speech Synthesis is most widely used statistical method of speech synthesis system. It is the technique generally used among the many speech researchers. This technique can also be combined with other advanced techniques like ANN to improve the performance. The HMM algorithms are basically inspired by mathematical model known as a Markov chain. HMM uses a probabilistic method uses pattern matching techniques where the observations are considered as a stochastic process outputs in association with Markov chain and finite set of probability distribution outputs. The HMM can also be modeled as left-right HMM. The name left right is derived due to its behavior and its topology where it can be modeled as the temporal flow of speech signal over time. [48] [49]

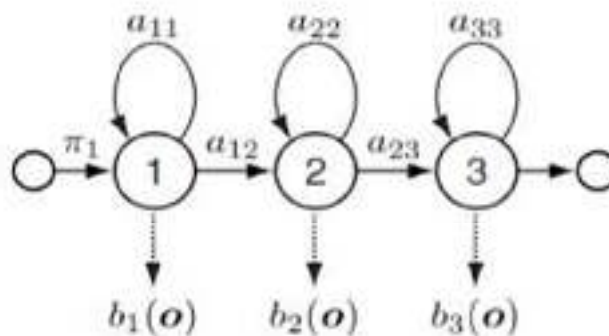


Figure 3.1 left-right Markov models [49]

Shows a 3-state left-to-right model, in which the state index increases or stays the same as time increments. In speech processing, the left-to-right models are often used as speech units to model speech parameter arrangements since they can suitably model signals whose properties progressively alter. [49]

3.4 Mel-Cepstrum Analysis

In sound handling the Mel-frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a direct cosine change of a log power spectrum on a nonlinear Mel scale of frequency. Cepstrum is characterized as the converse Fourier change of the logarithm of the spectrum, and excitation offers the advantages of low spectral distortion, low sensitivity to noise and efficiency in representing log spectral envelop. [50] In Mel-Cepstral examination the log spectrum is non-uniform dispersed in recurrence scale. [50] Mel-Cepstral coefficients can be determined from LPC coefficients but with non-linear exchange function. Not at all like the LPC model cares as it were channel poles, the Mel-Cepstral model incorporates

Mel Scale Approximation

Mel cepstral analysis uses logarithmic spectrum on Mel frequency scale to represent spectral envelopes and provide extra accuracy [48] the Mel frequency scale has a typical that it will expand the low frequency part and squeeze the high frequency part of the signal. Human ears have non-linear perception of frequency of sound, and are more sensitive to low frequency than to high frequency. Therefore, Mel frequency scale is more effective than linear frequency scale.

3.5 Linear Prediction Model

Linear prediction is a good tool for analysis of speech signals. Linear prediction models the human vocal tract as an infinite impulse response (IIR) system that produces the speech signal. For vowel sounds and other voiced regions of speech, which have a resonant structure and high degree of similarity overtime shifts that are multiples of their pitch period, this modeling produces an efficient representation of the sound. The figure below shows how the resonant structure of a vowel could be captured by an IIR system [51]

3.6 Fundamental Frequency Modeling

The natural frequency or fundamental frequency, often referred to simply as the fundamental, is defined as the lowest frequency of a periodic waveform. In music, the fundamental is the musical pitch of a note that is perceived as the lowest partial present. In terms of a superposition of sinusoids, the fundamental frequency is the lowest frequency sinusoidal in the sum. In some contexts, the fundamental is usually abbreviated as f_0 (or FF), indicating the lowest frequency counting from zero. In other contexts, it is more common to abbreviate it as

f_1 , the first harmonic. (The second harmonic is then $f_2 = 2 \cdot f_1$, etc. In this context, the zeroth harmonic would be 0 Hz.)

3.7 HMM state duration Modeling

A new method is proposed for modeling state duration in hidden Markov model (HMM) speech recognition systems. State transition probabilities are expressed conditional on how long the current state has been occupied. The conventional fixed-state transition probabilities a_{ij} are replaced by duration-dependent variables $a_{ij}(d)$ that depend on the time d already spent in state i . In this way, state transition and state duration probabilities are combined to form duration-dependent transition probabilities. The transition probabilities are derived from the cumulative density function (CDF) of state duration. The training of HMMs with duration-dependent transitions is based on maximum likelihood segmentation of training data, using the Viterbi algorithm. At each training iteration, the current HMM parameters are used to segment every training example. All the segments associated with each state are then used to update state observation and transition parameters. In experiments with a data set of spoken English alphabet, durational modeling improves the recognition accuracy by 5.6%.

3.8 Linear Prediction Model

Linear prediction is a good tool for analysis of speech signals. Linear prediction models the human vocal tract as an infinite impulse response (IIR) system that produces the speech signal. For vowel sounds and other voiced regions of speech, which have a resonant structure and high degree of similarity overtime shifts that are multiples of their pitch period, this modeling produces an efficient representation of the sound. The figure below shows how the resonant structure of a vowel could be captured by an IIR system [23].

3.9 CLUSTERGEN synthesizer

The clustergen synthesiser is a method for training models and using these models at synthesis time within the Festival Speech Synthesis System. The training requires well-recorded utterances and text transcriptions of what has been said. The best databases are those that are phonetically balanced. For our experiments, we have used the freely available CMU ARCTIC databases so that these experiments may be easily duplicated by others. Clustering is done by the Edinburgh Speech Tools CART tree-builder waggon. It has been extended to support vector predicates. CART trees are built in the normal way with waggon to find questions that split the data to minimize impurity. A tree is built for all the vectors

labelled with the same HMM state name. The impurity is calculated as follows: where N is the number of samples in the cluster and σ is the standard deviation for the MFCC feature over all samples in the cluster.

3.10 Decision Tree Building for Context Clustering

Decision tree based context clustering there are many contextual factors like phone identity factors, stress related factors, location factors that affect spectrum, F0 pattern and duration. To capture these effects, context dependent HMMs are used. However, as contextual factors increase, their combinations also increase exponentially. Therefore, model parameters cannot be estimated accurately with limited training data. Furthermore, it is impossible to prepare speech database which includes all combinations of contextual factors. To overcome this problem, a decision tree based context clustering technique is applied to distributions for spectrum, F0 and state duration in the same manner as HMM based speech recognition [52]. The decision tree based context clustering algorithms have been extended for MSD-HMMs. Since each of spectrum, F0 and duration has its own influential contextual factors, they are clustered independently as shown in figure 3. State durations of each HMM are modeled by a n -dimensional Gaussian and context dependent n -dimensional Gaussians are clustered by a decision tree. The spectrum part and F0 part of state output vector are modeled by multivariate Gaussian distributions and multi space probability distributions, respectively [52].

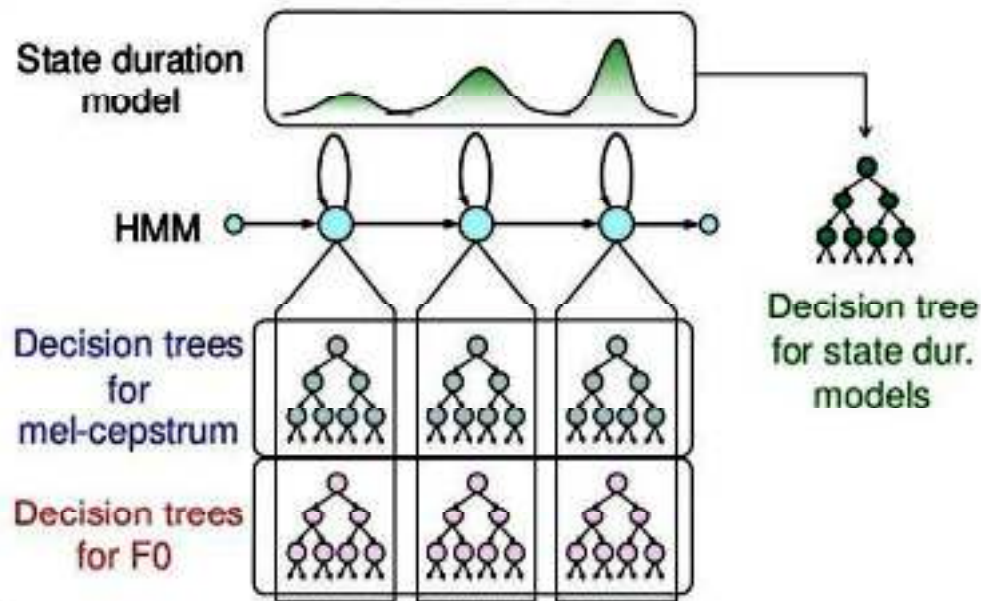


Figure 3.2 Decision Tree Building for Context Clustering [52]

HMM-based Speech Synthesis

The HMM-based speech synthesis is a statistical parametric model that extracts speech parameters from the speech database, trains them and produces the sound equivalent of the input text [50]. This method has the advantage of being able to synthesize speech with various speaker characteristics, speaking styles, emotions, and still produces reasonably natural sounding synthetic speech [53]. The decision tree clustered context-dependent HMMs are utilized for modeling the time varying speech parameters, and the SPS is sometimes called HMM based speech synthesis [36]. Unlike the unit selection method, SPS is able to generate speech that is not included in the original corpus by predicting the parameter values for a new context. It is also flexible in the sense that can be adapted to a different voice quality, speaking style, or speaker identity by using a small amount of corresponding speech material. In addition, it does not require as large a speech database as the unit selection methods, and the footprint is very small.

However, due to the parametric representation of speech, SPS suffers from lower segmental speech quality than unit selection synthesis [53].

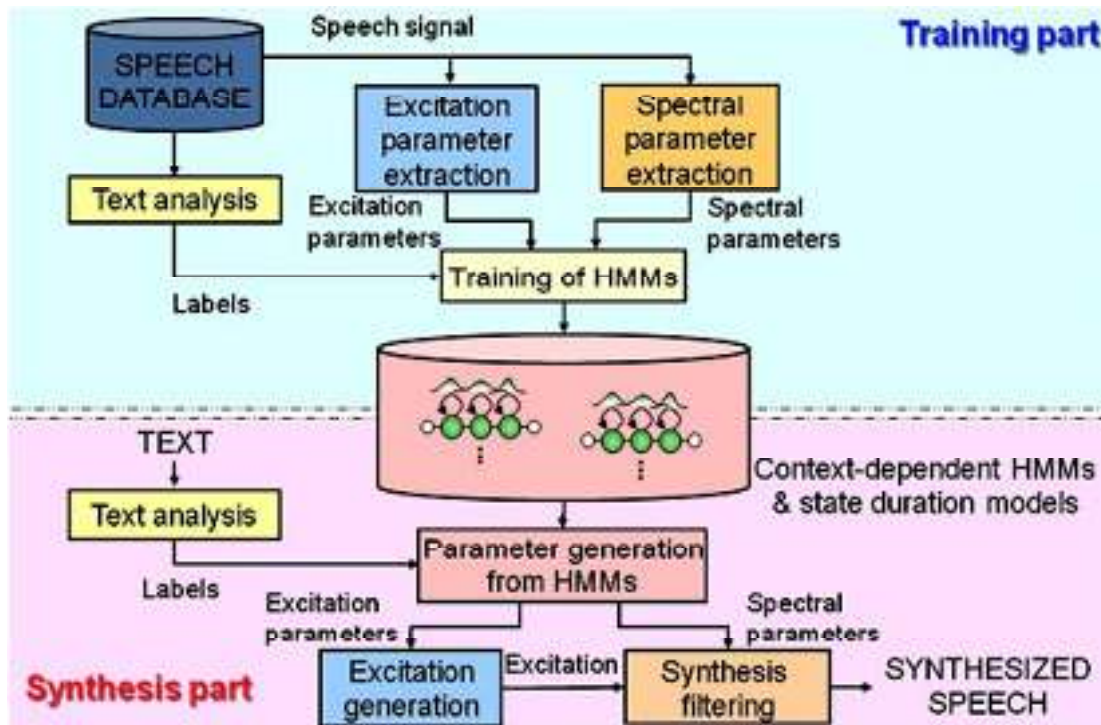


Figure 3.3 HMM-based speech synthesis system [53]

3.11 Text to speech synthesis system architecture

3.11.1 Introduction

This chapter deals with the Kambatta text to speech synthesis system architecture. It explains the whole process of the design, the algorithms used, their relation and interaction, the representation and description of components. The proposed architecture of speech synthesizer for Kambatta language by using statistical parametric approach is illustrated in Figure 3.4

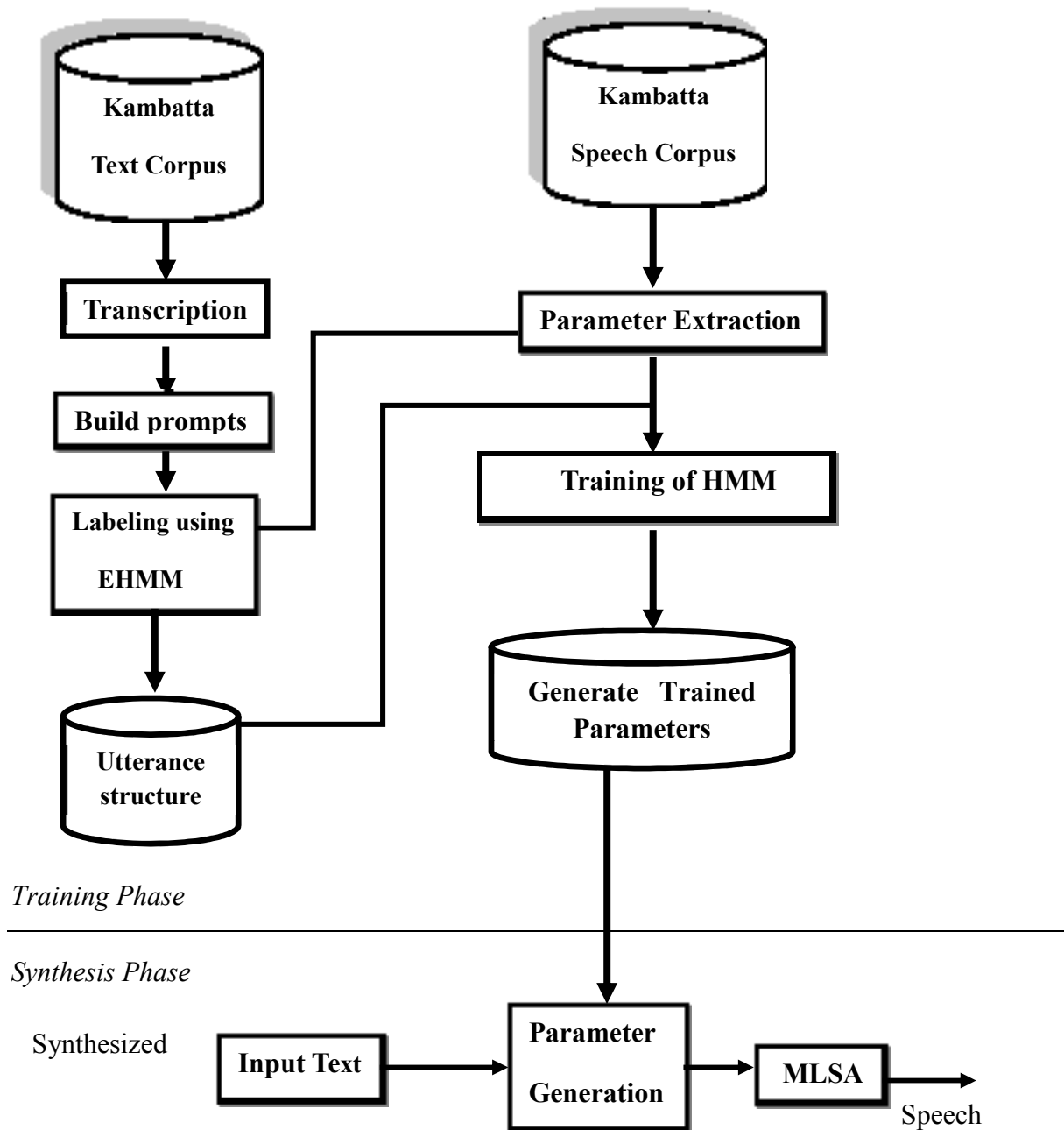


Figure 3.4 Proposed architecture of speech synthesizer for Kambatta language by using statistical parametric approach

3.11.2. Training Phase

In the training phase, the speech and text corpus are collected and labeled according language structure further processing. The speech parameters including spectral and excitation parameters are automatically extracted from the manually labeled speech corpus using speech

tools SPTK. After that the extracted speech parameters are automatically aligned with the manually labeled text to extract the utterance structure using ergodic hidden Markov model (EHMM) labeler. Next, from the labeled text and extracted parameters, the context dependent HMM-based training module are extracted and trained using the decision tree based context clustering methods. Finally, using expectation maximization (EM) algorithm, a trained voice of context-dependent HMMs and duration models are generated.

Transcription

The collected sentences corpus were then structuring suitable format for festival by transcribed, then put into a festvox file called data.txt.done which is used for creating prompts and the file would be copied to the path wcu_kk_abe/etc. An example of the file data.txt.done can be found in Appendix E

Labeling text and speech corpus

Labeling is the process of giving a label for each speech signal in the utterance or generates the labeled utterances. The manually labeled text and extracted speech corpus are aligning using text labeler EHMM to generate the utterance for each sentence. In this research, the labeled utterances are generated by speech tools using the festival to understand Kambatta letter to sound transliteration tools using automatic EHMM labeler. For example, from the given texts “wcu_kk_abe_002 "zaba bao booru zakko fanqalanobaa jaata luushsho xawu hinaten zumanobaa.", and generate the utterance structure parameters in the form of (utt) (see in Appendix A).

A **phone set**; is a set of symbols which may be further defined in terms of features extraction such as vowel, place of articulation, manner of articulation for constants and type of vowel in which the set of features and their values must be defined. In Kambatta, there are around 25 phone sets and each has its own standard phone features including manner of articulation, height of vowel, vowel length, lip rounding, consonant sound and vowel frontless. Therefore, the concept of phone sets is very crucial to a number of different subsystems within festival. The festival also supports multiple phone sets simultaneously and allows mapping between sets when necessary. As a result, the Kambaata letter to sound transliteration, waveform 25 synthesized, all require the definition of the phone set must declared before they will operate (see Appendix B).

Parameterization

In this work, mainly we have two parameters the spectrum and excitation parameters. In order to produce natural sounding speech at low bit rate, the parameters representing speech information effectively need to be extracted from source files. The excitation signal usually requests fundamental frequency F_0 , spectral envelopes information and voiced and unvoiced decision parameters. The spectrum parameters consist of Mel-frequency Cepstrum Coefficients (MFCC) and are used as spectrum parameters and the fundamental frequency is used as a source of speech.

In this thesis work, before going to extract the parameter features from the raw speech corpus, first we have to prune the given data to erase noise and generate the labeled speech from the raw speech database manually. The extracted parameters: the spectrum and excitation consists of two streams. The first stream contains the spectrum part along with the delta and delta of delta values and the second stream includes the logarithm of the fundamental frequency ($\log F_0$) along with their delta, and delta of delta values respectively. Each stream of information is modeled separately and delta and delta of delta values are used to model the dynamic nature of the speech. Finally, using the speech signal parametric toolkits (SPTK) tools parameters features vectors are extracted.

Text Analysis

From the whole Text-to-Speech system design the first is text analysis, which identifies pronounceable words from raw text. In the current text analysis field, text normalization is considered as a key component of text analysis in TTS. It is frequently used when converting text to speech. Like Numbers, dates, acronyms, and abbreviations are non-standard words that need to be pronounced differently depending on context. Non-standard words (NSW) cannot be noticed by an application of letter-to-sound and may be recognized as different standard words depending on both the local text and the text type. For example “6” would be pronounced as “lihuu”. Because of this all of non-standard representations should normalize, or in other words convert to standard words. Initially text tokenization methodologies are implemented for the language before normalization is done. White-space is most commonly used delimiter between words and is extensively used for tokenization []. As a delimiter for the tokenization process, whitespaces, tabs, new lines and carriage return characters are used to tokenize Kambatta text. For example, the Kambatta sentence “Kanniichch biree bashilat

woraqattat maassa” is tokenized as “Kanniichch” “biree” “bashilat” “woraqattat” and “maassa” tokens. After that Festival converts the give sentence into ordered list of tokens for further process. Once the text has been tokenized, text normalization is carried out.

Phonetic Analysis

After text analyzed to pronounceable word, the second phase of natural language processing module of text to speech system begin. This phase converts the analyzed text word to its pronunciation form. This TTS system is phonetic analysis. Phonetic analysis also known as word analysis, phonics or decoding is the process of using the relationships between spelling and pronunciation at the letter, syllable, and word levels to figure out unfamiliar words. [54]

The phonetic examination module takes the normalized word strings from the content preparing module and produces elocution for each word. The articulation is given not fair as a list of phones, but too a syllabic structure and lexical push. The strategy for finding the articulation of a word is by letter to sound rules. For creating Kambatta synthesizer, we utilized both dictionaries and hand composed letter to sound run the show sets. Hand composed letter to sound rules are setting subordinate re-write rules which are connected in arrangement mapping strings of letters to strings of phones.

The festival system gives a dictionary subsystem upon which lexical examination might be performed for the language. Utilizing this system, a compiled dictionary is ready utilizing the words aiming for building the synthesizer model (training set). Festival moreover gives an elective way of speaking to words pronunciation called the Adenda. But, since the Adenda is looked directly when a modern instance is experienced, it is less proficient than the compiled dictionary which takes after a dual look instrument.

In addition to the lexicon entries prepared explicitly for the system, a letter to sound rule is implemented as a mechanism to handle words whose pronunciation is not given in the lexicon.

Chapter Four

Results and Discussion

4.1 The Development Environment and Tools

In order to set up a workable atmosphere for introducing the current research demo, a few programmer installations were necessary. On a typical laptop with 1.9 GB of memory and an Intel(R) Celeron (R) CPU N2830 @ 2.16GHz*2, the experimental environment was set up. Ubuntu 14.04 LTS was the operating system being used. For this research demo to be successful, a number of software programmers have to be downloaded and installed.

- Speech tools - It is free software based on library of C++ functions for the speech processing used for reading, writing, converting and supporting speech processing objects.
- Festival - It is a multilingual TTS system which provides a platform for developing TTS systems.
- Festvox - open source software based on festival for voice building process includes program tool such as phone set, lexicon, phrasing and etc.
- SPTK-3.9- It is freely available software for speech signal processing and extracting speech parameters and to re-synthesize speech from the parameters
- Adobe Audition - It is software used for sound visualization, recording and management
- Laptop computer with Ubuntu operating system, Intel core i5 with 2.5 GHz processor speed, 4.00 GB RAM and 465 GB hard disk capacity
- Microsoft office 2010 for documentation
- Microphone for recording speech data.

4.2 Preparing Questionnaire

A questionnaire is created before the evaluation and contains questions regarding how natural and understandable the synthesised speech is. The purpose of the first question is to assess how understandable the synthesised speech is, and the purpose of the second question is to assess how closely the synthesised speech resembles actual human speech. Both testing methods employ the same set of questions (see Appendix D).

4.3 Testing and Evaluations

In this study, the naturalness and understandability of the synthesised speech are assessed using two measurements: the Mean Opinion Score (MOS) and the Mel Cepstral Distortion (MCD).

The mean opinion score (MOS) test is selected for this evaluation, which allowed us to score and compare the quality of TTS systems with respect to naturalness and intelligibility.

Table 4.1 Scales used in MOS

MOS value	Quality
5	Excellent
4	Very Good
3	Good
2	fair
1	bad

The performance of the systems will be evaluated by the evaluators. The raters then provide their rankings in accordance with the MOS scale. In each case, a questionnaire is used to elicit the evaluator's viewpoint. The survey is appended as Appendix D. The conclusions of the evaluation are reported in the format outlined in Appendix E.

Four hundred fifty sentences are used for training and five sentences for testing. Six speakers of the language are tested the system. The language speakers were made aware of the special nature of synthetic speech produced and generally the text to speech systems so that they could appropriately adjust their prospect. After they listen the synthesized speech from the system, the invited Kambaata speakers are given with the questionnaire to evaluate intelligibility and naturalness of the synthesized speech.

Table 4.2 MOS score of the sound

Test Data(Sentence)	Naturalness	Intelligibility
1	2.41	3.28
2	2.39	3.03
3	2.55	2.78
4	2.45	2.59
5	2.7	3.2
Average Score	2.5	2.97

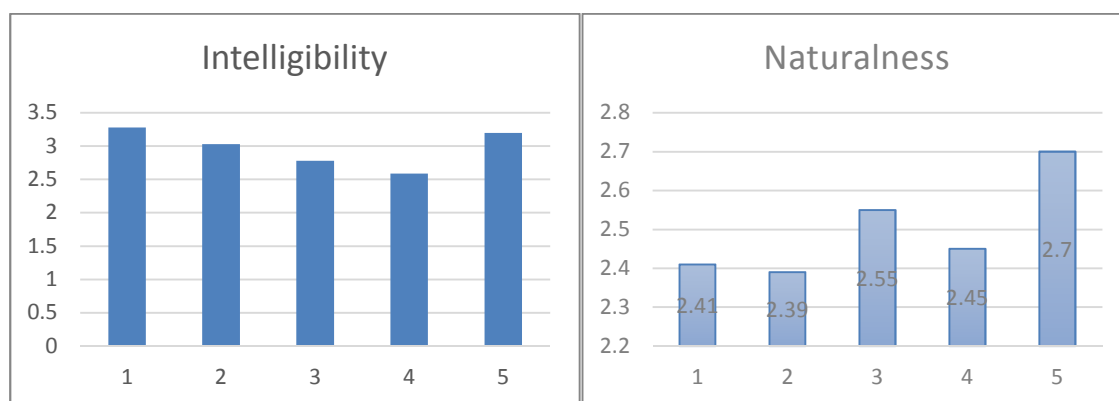


Figure 4.1 MOS Score of the Sound in graph

Finally, according to mean opinion score, the mean and standard deviation cumulative results are calculated as per the respondents' responses

The system compares and contrasts the advantageous characteristics of synthetic and natural sound. Ten-fold cross validation rule, a popular and efficient training approach, forms the basis of our system's operation. The result achieved using Mel cepstral distortion (MCD) is 7.2, which is good.

Chapter Five

Conclusions and Recommendation

5.1 Conclusions

Speech synthesis, as is generally known, serves a number of purposes in modern human activities, including helping the disabled and working in the telecommunications sector. The system entails text collection, text preprocessing, phonetically matched sentences, sentence recording, annotated speech database creation, and prototype design. Phases of training and synthesis are included in architecture. First, the text and speech corpus are manually produced for processing during the training phase. Mel-Cepstral analysis is used to get the excitation and spectrum parameters (coefficients of Mel-Cepstral analysis) from the speech database. The text corpus and speech parameters are then aligned to construct an utterance using an automatic EHMM labeler.

The f_0 , spectrum, and duration are trained to create the context dependent HMM parameters model based on decision tree clustering approach or classification and regression tree in the synthesis phase after an arbitrarily given text to be preprocessed is first converted into a context-based label sequence. Then, the context dependent HMMs' training speech parameters (spectral and excitation) were matched to these attributes. In conclusion, artificial speech is produced using the MLSA speech parameter algorithms. In this work, a statistical parametric technique is used to create a speech synthesiser for the Kambaata language. However, because it takes a lot of time and requires in-depth linguistic understanding, not every aspect of the Kambaata language was taken into account. The evaluation method of mean opinion score was applied.

5.2 Recommendations

The Kambaata Speech Synthesis System is built in this thesis work using a hidden markov model (HMM). However developed systems performance is good; except there are Features that are not being included in this thesis work. Therefore to have a speech synthesizer that considers all speech features, the researcher recommends extensions of work that increase the quality of the synthesized speech.

In the future, it is better to apply the deep learning methods or deep neural network methods (DNN) which will lead to hybrid techniques that incorporate both the benefits of statistical parametric speech based on HMM and unit selection synthesis method.

This study did not consider non-standard words such as numbers, dates, abbreviations, etc. which are challenging in designing the speech synthesis unless there is a standard corpus. This is basically because of the nonexistence of specialized linguistic resources that can be used for developing letter to sound rule and compiled lexicon.

Reference

- [1] R. C. & J. Glass, "Vowel Classification Based on Analysis-by-Synthesis," canada, 2004.
- [2] M. Takashi, HMM-Based Speech Synthesis and Its Applications, tokiyo: Unpublished Masters thesis, 2002.
- [3] A. J. a. P. Mythili, Developing a Child Friendly Text-to-Speech System, India: HindawiPublishing Corporation, 2008.
- [4] T. Raitio, Voice source modelling techniques for statistical parametric speech synthesis, Unpublished Doctoral Dissertations, 2015.
- [5] A. W. Black, CLUSTERGEN: A Statistical Parametric Synthesizer using, Pittsburgh, PA, USA, 2006.
- [6] A. W. a. L. A. Black, Building Synthetic Voices, 2003.
- [7] L. S., Review of Speech Synthesis Technology, finland: MSc.Thesis. Laboratory of Acoustics and Audio Signal Processing, HelsinkiUniversity of Technology, 1999.
- [8] TewodrosAbebe, Text-to-Speech Synthesizer for WolayttaLanguage, Addis Ababa, Ethiopia, 2009.
- [9] Speech synthesis, 2010.
- [10] T. Dutoit, A Short Introduction to Text-to-Speech, boston: KluwerAcademic Publishers, 1997.
- [11] S. Lemmetty, Review of Speech Synthesis Technology, 1999.
- [12] M. MoutranAssaf, A Prototype of an Arabic Diphone Speech Synthesizer in Festival, 2005.
- [13] A. W. A. a. L. D. Azhar Ali Shah, "Bi-Lingual Text to Speech," 2004.
- [14] D. H. Arficho, Ideophones in Kambaata, addis ababa, 2020.
- [15] H. Z. A. W. B. K. Tokuda, An HMM based speech synthesis system applied, Nagoya, 2002.
- [16] A. H. M. S. Sacha K., An HMM-Based Speech Synthesis System applied to German, berlin, 2002.

- [17] B. Ntsako, Text- To-Speech Synthesis System for Xitsonga using Hidden Markov, 2015.
- [18] A. Tafere, A GENERALIZED APPROACH TO AMHARIC TEXT-TO-SPEECH (TTS), addis ababa: ADDIS ABABA UNIVERSITY, 2010.
- [19] S. Tadesse, Concatenative Text-To-Speech System for Afaan Oromo Language, addis ababa, 2011.
- [20] A. Kiflu, Unit Selection Based Text-to-Speech Synthesizer for Tigrinya Language, addis ababa, 2004.
- [21] HenockLulseged, Concatenative Text-to-Speech (TTS) synthesis for Amharic language, addis ababa, 2003.
- [22] E. Donovan, Trainable Speech Synthesis, united kingdom, 1996.
- [23] HabtamuHaye, "Amharic concatenative Text- To-Speech (TTS) synthesis system using," addis ababa, 2007.
- [24] F. P. W., Cummings Otolaryngology-Head & Neck Surgery, Philadelphia, 2015.
- [25] R. C. a. S. C., Operative Techniques in Laryngology., Leipzig, 2008.
- [26] V.-H. A. e. al and D. C. a. colleagues, Dynamics of Intrinsic Laryngeal Muscle Contraction., 2018.
- [27] T. I. R., Fascinations with the Human Voice., 2010.
- [28] Steven K Smith, Digital Signal Processing, A Practical Guide for Engineers and, USA, 2003.
- [29] T. & K. E. Styger, Formant Synthesis. In E. Keller (ed.), Fundamentals of Speech, Chichester, 1994.
- [30] J. Allen, M. Sharon and Dennis, The MITalk system, Cambridge, 1987.
- [31] P. Rubin, T. Baer and P. Mermelstein, "An articulatory synthesizer for perceptual research".
- [32] B. Carr, G. F. R. Ellis, G. Gibbons, J. B. Hartle, T. Hertog and Roger, Stephen William Hawking CH CBE. 8 January 1942—14 March 2018, 2019.
- [33] T. Dutoit, A Short Introduction to Text-to-Speech, Boston: KluwerAcademic Publishers, 1997.

- [34] H. J. a. A. M. Medina A., Towards the Speech Synthesis of Raramuri: A Unit Selection Approach based on Unsupervised Extraction of Suffix Sequences, A. Gelbukh.
- [35] T. Styger and E. Keller, Formant Synthesis. In E. Keller (ed.), Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges, 1994.
- [36] J. D. Pawar and G. A. Naik, Implementation of a TTS System for Devanagari Konkani Language using Festiva, Goa, India, 2017.
- [37] 2. B. C. T .SRAVYA & R.VASAVI, SPEECH SYNTHESIS, 2010.
- [38] H. Lulseged, Concatinative Text-to-Speech (TTS) synthesis for Amharic language, addis ababa, 2003.
- [39] Habtamu Taye, Diphone based Text-to-Speech Synthesis System for Amharic, addis ababa, 2007.
- [40] L. H. a. M. M, Developing Tigrigna Text to Speech Synthesizer, addis ababa, 2010.
- [41] Nadew Tademe, Formant based speech synthesis for Amharic vowels, addis ababa, 2008.
- [42] TesfayYihdego, Diaphone Based Text-To-Speech Synthesis System for Tigrigna, addis ababa, 2004.
- [43] B. Kasaye, DEVELOPING A SPEECH SYNTHESIZER FOR AMHARIC, addis ababa, 2008.
- [44] Grimes, web edition of Ethnologue, 2000.
- [45] Y. Treis, A Grammar of Kambaata (Ethiopia), germny, 2008.
- [46] H. T. T. N. M. a. T. K. Zen, Nitech HMM-based speech synthesissystem for the Blizzard Challenge, 2005.
- [47] d. datcu, automatic recogintion of facial expration, 2004.
- [48] V. M. T. I. Oliveira L., A Rule Based Text-to-Speech System forPortuguese, 1992.
- [49] K. T. T. K. Heiga Zen, An Introduction of Trajectory Modelin to HMM-Based Speech Synthesis, japan, 2002.
- [50] W. L. a. H. Mark, Speech Synthesis using Mel-CepstralCoeffiecent Feature, urbana: Unpublished Senior thesisin Electrical Engineering University of Illinois at

UrbanaChampaign, 2015.

- [51] S. K. D. M. S. Mukherjee, A BENGALI HMM BASED SPEECH SYNTHESIS SYSTEM, Bengali: Unpublished Masters Thesis, 2012.
- [52] K. M. T. M. N. a. K. T. Tokuda, Multi-space probability, 2002.
- [53] H. A. Dr, Statistical Parameter Speech based on HMM speech synthesis (HTS), Gujarati, 2006.
- [54] T. Paul, Text to Speech Synthesis, Edinburgh, 2007.
- [55] T. M. S. D. Sadhana Gopal, "A simple phoneme based speech recognition system," Vols. Volume-2, no. Issue-4, 2014.
- [56] D. H. Arficho, Ideophones in Kambaata, addis abab, 2020.

Appendix

Appendix A: Utterance Structure

(EST_File utterance

DataType ascii

Version 2

EST_Header_End

Features max_id 241 ; type Text ; iform "\ zahishsh dagumbu zahan galano qalbus
foolooccumbu mexxurraa zumaa karooriinn.\""; filename prompt-utt/uniph_053.utt ; fileid
uniph_053 ;

Stream_Items

1 id _1; name zahishsh ; whitespace "" ; prepunctuation "" ;
2 id _2; name dagumbu ; whitespace " " ; prepunctuation "" ;
3 id _3; name zahan ; whitespace " " ; prepunctuation "" ;
4 id _4 ; name galano ; whitespace " " ; prepunctuation "" ;
5 id _5 ; name qalbus ; whitespace " " ; prepunctuation "" ;
6 id _6 ; name qalbus ; whitespace " " ; prepunctuation "" ;
7 id _7 ; name foołooccumbu ; whitespace " " ; prepunctuation "" ;
8 id _8 ; name mexxurraa ; whitespace " " ; prepunctuation "" ;
9 id _9 ; name zumaa ; whitespace " " ; prepunctuation "" ;
10 id _10 ; name karooriinn ; punc . ; whitespace " " ; prepunctuation "" ;
11 id _20 ; name karooriinn ; pbreak B ; pos nil ;
12 id _21 ; name . ; pbreak B ; pos punc ;
13 id _19 ; name zumaa ; pbreak NB ; pos nil ;
14 id _18 ; name mexxurraa ; pbreak NB ; pos nil ;

15 id _17 ; name foolooccumbu ; pbreak NB ; pos nil ;

16 id _16 ; name qalbus ; pbreak NB ; pos nil ;

17 id _15 ; name galano ; pbreak NB ; pos nil ;

18 id _14 ; name zahan ; pbreak NB ; pos nil ;

19 id _13 ; name dagumbu ; pbreak NB ; pos nil ;

20 id _12 ; name zahishsh ; pbreak NB ; pos nil ;

21 id _11 ; name dagumbu ; pbreak NB ; pos nil ;

22 id _23 ; name syl ; stress 1 ;

23 id _25 ; name syl ; stress 0 ;

24 id _28 ; name syl ; stress 0 ;

25 id _30 ; name syl ; stress 1 ;

26 id _33 ; name syl ; stress 0 ;

27 id _36 ; name syl ; stress 1 ;

28 id _38 ; name syl ; stress 0 ;

29 id _41 ; name syl ; stress 1 ;

30 id _44 ; name syl ; stress 0 ;

31 id _46 ; name syl ; stress 1 ;

32 id _49 ; name syl ; stress 0 ;

33 id _51 ; name syl ; stress 1 ;

34 id _54 ; name syl ; stress 0 ;

35 id _57 ; name syl ; stress 1 ;

36 id _59 ; name syl ; stress 1 ;

37 id _61 ; name syl ; stress 0 ;

38 id _64 ; name syl ; stress 1 ;
39 id _67 ; name syl ; stress 0 ;
40 id _70 ; name pau ; dur_factor 1 ; end 0.0435 ;
41 id _24 ; name m ; dur_factor 1 ; end 0.4542 ;
42 id _26 ; name d ; dur_factor 1 ; end 0.4548 ;
43 id _27 ; name ee ; dur_factor 1 ; end 0.6025 ;
44 id _29 ; name n ; dur_factor 1 ; end 0.795 ;
45 id _31 ; name b ; dur_factor 1 ; end 1.85 ;
46 id _32 ; name aa ; dur_factor 1 ; end 1.84 ;
47 id _34 ; name j ; dur_factor 1 ; end 1.505 ;
48 id _35 ; name a ; dur_factor 1 ; end 1.66 ;
49 id _37 ; name b ; dur_factor 1 ; end 1.905 ;
50 id _39 ; name y ; dur_factor 1 ; end 2.27 ;
51 id _40 ; name a ; dur_factor 1 ; end 2.365 ;
52 id _42 ; name k ; dur_factor 1 ; end 2.695 ;
53 id _43 ; name ee ; dur_factor 1 ; end 3.45 ;
54 id _45 ; name aa ; dur_factor 1 ; end 3.005
55 id _47 ; name f ; dur_factor 1 ; end 4.875 ;
56 id _48 ; name i ; dur_factor 1 ; end 4.13 ;
57 id _50 ; name aa ; dur_factor 1 ; end 4.64 ;
58 id _52 ; name r ; dur_factor 1 ; end 5.115 ;
59 id _53 ; name aa ; dur_factor 1 ; end 5.441 ;
60 id _55 ; name k ; dur_factor 1 ; end 5.78 ;

61 id _56 ; name a ; dur_factor 1 ; end 5.512 ;
62 id _58 ; name t ; dur_factor 1 ; end 6.01 ;
63 id _60 ; name ; dur_factor 1 ; end 6.580 ;
64 id _62 ; name * ; dur_factor 1 ; end 6.3 ;
65 id _63 ; name e ; dur_factor 1 ; end 6.755 ;
66 id _65 ; name h ; dur_factor 1 ; end 6.94 ;
67 id _66 ; name ee ; dur_factor 1 ; end 6.955 ;
68 id _68 ; name r ; dur_factor 1.5 ; end 7.02 ;
69 id _69 ; name a ; dur_factor 1.5 ; end 7.04 ;
70 id _71 ; name pau ; dur_factor 1 ; end 7.34 ;
71 id _72 ; name Accented ;
72 id _73 ; name Accented ;
73 id _74 ; name Accented ;
74 id _75 ; name Accented ;
75 id _76 ; name Accented ;
76 id _77 ; name Accented ;
77 id _102 ; f0 110 ; pos 3.41 ;
78 id _100 ; f0 121.75 ; pos 3.07 ;
79 id _101 ; f0 112.5 ; pos 3.08 ;
80 id _99 ; f0 123.75 ; pos 2.87 ;
81 id _98 ; f0 113.75 ; pos 2.86 ;
82 id _95 ; f0 126.25 ; pos 2.43 ;
83 id _96 ; f0 124.25 ; pos 2.52 ;

84 id _97 ; f0 115.625 ; pos 2.53 ;
 85 id _94 ; f0 116.25 ; pos 2.42 ;
 86 id _92 ; f0 130.5 ; pos 1.53 ;
 87 id _93 ; f0 121.25 ; pos 1.54 ;
 88 id _91 ; f0 132.5 ; pos 1.33 ;
 89 id _90 ; f0 122.5 ; pos 1.32 ;
 90 id _87 ; f0 134.375 ; pos 1 ;
 91 id _88 ; f0 132.375 ; pos 1.09 ;
 92 id _89 ; f0 123.75 ; pos 1.1 ;
 93 id _86 ; f0 124.375 ; pos 0.99 ;
 94 id _84 ; f0 134.875 ; pos 0.76 ;
 95 id _85 ; f0 125.625 ; pos 0.77 ;
 96 id _83 ; f0 136.875 ; pos 0.56 ;
 97 id _82 ; f0 126.875 ; pos 0.55 ;
 98 id _79 ; f0 139.375 ; pos 0.12 ;
 99 id _80 ; f0 137.375 ; pos 0.21 ;
 100 id _81 ; f0 128.75 ; pos 0.22 ;
 101 id _78 ; f0 129.375 ; pos 0.11 ;
 102 id _135 ; name BB ;

End_of_Stream_Items

Relations

Relation Token ; "(" ")" ;

11 21 1 0 0 0

1 1 0 11 2 0

12 20 2 0 0 0

2 2 0 12 3 1

13 19 3 0 0 0

3 3 0 13 4 2

14 18 4 0 0 0

4 4 0 14 5 3

15 17 5 0 0 0

5 5 0 15 6 4

16 16 6 0 0 0

6 6 0 16 7 5

17 15 7 0 0 0

7 7 0 17 8 6

18 14 8 0 0 0

8 8 0 18 9 7

19 13 9 0 0 0

9 9 0 19 10 8

20 11 10 0 21 0

21 12 0 0 0 20

10 10 0 20 0 9

End_of_Relation

Relation Word ; "(" ")" ;

1 21 0 0 2 0

2 20 0 0 3 1

3 19 0 0 4 2

4 18 0 0 5 3

5 17 0 0 6 4

6 16 0 0 7 5

7 15 0 0 8 6

8 14 0 0 9 7

9 13 0 0 10 8

10 11 0 0 0 9

End_of_Relation

Relation Segment ; "(" ")" ;

1 40 0 0 2 0

2 41 0 0 3 1

3 42 0 0 4 2

4 43 0 0 5 3

5 44 0 0 6 4

6 45 0 0 7 5

7 46 0 0 8 6

8 47 0 0 9 7

9 48 0 0 10 8

10 49 0 0 11 9

11 50 0 0 12 10

12 51 0 0 13 11

13 52 0 0 14 12

14 53 0 0 15 13

15 54 0 0 16 14

16 55 0 0 17 15

17 56 0 0 18 16

18 57 0 0 19 17

19 58 0 0 20 18

20 59 0 0 21 19

21 60 0 0 22 20

29 68 0 0 30 28

30 69 0 0 31 29

31 70 0 0 0 30

End_of_Relation

Relation SylStructure ; "(" ")" ;

15 41 12 0 0 0

12 22 1 15 13 0

16 42 13 0 17 0

17 43 0 0 0 16

13 23 0 16 14 12

18 44 14 0 0 0

14 24 0 18 0 13

1 21 0 12 2 0

21 45 19 0 22 0

22 46 0 0 0 21

19 25 2 21 20 0

23 47 20 0 24 0
24 48 0 0 0 23
20 26 0 23 0 19
2 20 0 19 3 1
27 49 25 0 0 0
25 27 3 27 26 0
28 50 26 0 29 0
29 51 0 0 0 28
26 28 0 28 0 25
3 19 0 25 4 25
32 52 30 0 33 0
33 53 0 0 0 32
30 29 4 32 31 0
34 54 31 0 0 0
31 30 0 34 0 30
4 18 0 30 5 3
36 55 35 0 37 0
37 56 0 0 0 36
35 31 5 36 0 0
5 17 0 35 6 4
39 57 38 0 0 0
38 32 6 39 0 0
6 16 0 38 7 5
41 58 40 0 42 0
42 59 0 0 0 41
40 33 7 41 0 0
7 15 0 40 8 6
44 60 43 0 45 0
45 61 0 0 0 44
43 34 8 44 0 0
8 14 0 43 9 7
49 62 46 0 0 0
46 35 9 49 47 0

50 63 47 0 0 0
47 36 0 50 48 46
51 64 48 0 52 0
52 65 0 0 0 51
48 37 0 51 0 47
9 13 0 46 10 8
55 66 53 0 56 0
56 67 0 0 0 55
53 38 10 55 54 0
57 68 54 0 58 0
58 69 0 0 0 57
54 39 0 57 0 53
10 11 0 53 11 9
11 12 0 0 0 10

End_of_Relation

Relation IntEvent ; "(" ")" ;

1 71 0 0 2 0
61

2 72 0 0 3 1
3 73 0 0 4 2
4 74 0 0 5 3
5 75 0 0 6 4
6 76 0 0 0 5

End_of_Relation

Relation Intonation ; "(" ")" ;

7 71 1 0 0 0
1 22 0 7 2 0
8 72 2 0 0 0
2 25 0 8 3 1
9 73 3 0 0 0
3 27 0 9 4 2
10 74 4 0 0 0
4 29 0 10 5 3

11 75 5 0 0 0

5 35 0 11 6 4

12 76 6 0 0 0

6 38 0 12 0 5

End_of_Relation

Relation Target ; "(" ")" ;

17 101 1 0 0 0

1 40 0 17 2 0

18 98 2 0 19 0

19 99 0 0 20 18

20 100 0 0 0 19

2 41 0 18 3 1

21 97 3 0 0 0

3 44 0 21 4

38 80 14 0 0 0

14 66 0 38 15 13

39 78 15 0 40 0

40 79 0 0 0 39

15 67 0 39 16 14

41 77 16 0 0 0

16 69 0 41 0 15

End_of_Relation

Relation Phrase ; ()

2 21 1 0 3 0

3 20 0 0 4 2

4 19 0 0 5 3

5 18 0 0 6 4

6 17 0 0 7 5

7 16 0 0 8 6

8 15 0 0 9 7

9 14 0 0 10 8

10 13 0 0 11 9

11 11 0 0 0 10

1 102 0 2 0 0

End_of_Relation

End_of_Relations

End_of_Utterance

Appendix B: Kambaata Phone sets

```
(defPhoneSet
Wcu_kembab
;Phone Features
(;vowel or consonant
(vc + -)
;vowel length: short long diphthong schwa
(vlng s l d a 0)
;vowel height: high mid low
(vheight 1 2 3 0 -)
;vowel frontness: front mid back
(vfront 1 2 3 0 -)
;;lip rounding
(vrnd + - 0)
;consonant type: stop fricative affricative nasal liquid
(ctype s f a n l r 0)
;place of articulation: labial alveolar palatal labio-dental
;dental velar
(cplace l a p b d v g 0)
;;consonant voicing
(cvox + - 0)
)
(
```

(pau - 0 - - - 0 0 -) ;; slience ...

(SIL - 0 0 0 0 0 0 -);;silence....

(a + s 3 3 - 0 0 0)

(e + s 2 1 - 0 0 0)

(i + s 1 1 - 0 0 0)

(o + s 2 3 + 0 0 0)

(u + s 1 3 + 0 0 0)

(aa + l 3 3 - 0 0 0)

(ee + l 2 1 - 0 0 0)

(ii + l 1 1 - 0 0 0)

(oo + l 2 3 + 0 0 0)

(uu + l 1 3 + 0 0 0)

(b - 0 0 0 0 s l +)

(c - 0 0 0 0 a p +)

(d - 0 0 0 0 s a +)

(f - 0 0 0 0 f b -)

(g - 0 0 0 0 s v +)

(h - 0 0 0 0 f g -)

(j - 0 0 0 0 a p +)

(k - 0 0 0 0 s v -)

(l - 0 0 0 0 l a +)

(m - 0 0 0 0 n l +)

(n - 0 0 0 0 n a +)

(q - 0 0 0 0 s v +)

(r - 0 0 0 0 r a +)

(s - 0 0 0 0 f a +)

(t - 0 0 0 0 s a -)

(w - 0 0 0 0 r l +)

(x - 0 0 0 0 s a +)

(y - 0 0 0 0 r p +)

(z - 0 0 0 0 f a +)

(ch - 0 0 0 0 a p -)

(ph - 0 0 0 + s l +)

(sh - 0 0 0 0 f p -)

)

)

Appendix C: Sample of Kambaata letter to sound rule transliteration

```
#include<iostream>
#include<fstream>
using namespace std;
int isVowel(char ch);
void word_to_pronunciation(string word, const char *output_filename);
ofstream outf;
ifstream inpf;
string normalize(string input);
string de_normalize(char input);
int main()
{
string word;
/*if(argc != 3)
{
cout<<"Error: Usage "<<argv[0]<<" word_input_filename
word_feature_output_filename"<<endl;
return 1;
}*/
inpf.open("/home/ak/abenis/tts/wcu_kk_abe/bin/word");
if(inpf.fail())
{
cout<<"unable to open file"<<endl;
return 1;
}
inpf>>word;
word_to_pronunciation(word, "/home/ ak/abenis/tts/wcu_kk_abe/wordpronunciation");
inpf.close();
return 0;
}
void word_to_pronunciation(string word, const char *output_filename)
{
outf.open(output_filename);
```

```

if(outf.fail())
66
{
cout<<"unable to open file"<<output_filename<<endl;
return;
}
string normalized_word = normalize(word);
cout<<normalized_word<<endl;
int len = normalized_word.length();
outf<<"(set! wordstruct '( ( (";
int index = 0;
while(index < len-2)
{
if(isVowel(normalized_word[index+1]))
{
if(index == 0) outf<<de_normalize(normalized_word[index])<<"
"<<de_normalize(normalized_word[index+1])<<" ) 1 ) ( ( ";
else outf<<de_normalize(normalized_word[index])<<"
"<<de_normalize(normalized_word[index+1])<<" ) 0 ) ( ( ";
index = index + 2;
}
else
{
if(index == 0) outf<<de_normalize(normalized_word[index])<<" ) 1 )
( ( ";
else outf<<de_normalize(normalized_word[index])<<" ) 0 ) ( ( ";
index++;
}
}
if(index + 1 == len)
outf<<de_normalize(normalized_word[index])<<" ) 0 ) )"<<endl;
else
outf<<de_normalize(normalized_word[index])<<"

```

```

"<<de_normalize(normalized_word[index+1])<<" ) 0) ))"<<endl;
outf.close();
}
string de_normalize(char input)
{
string output;
string temp = "x";
switch(input){
case 'A': {output = "aa";break;}
case 'E': {output = "ee";break;}
case 'I': {output = "ii";break;}
67
case 'O': {output = "oo";break;}
case 'U': {output = "uu";break;}
case 'C': {output = "ch";break;}
case 'D': {output = "dh";break;}
case 'P': {output = "ph";break;}
case 'S': {output = "sh";break;}
case 'N': {output = "ny";break;}
case 'Z': {output = "zy";break;}
case 'T': {output = "ts";break;}
case 'J': {output = ""};break;}
default: {temp[0] = input; output = temp;}
}
return output;
}
string normalize(string input)
{
string output;
int i = 0;
int len = input.length();
while (i < len-1)
{

```

```

if(input[i] == 'a')
{
switch(input[i+1]){
case 'a': {output =output + "A";break; }
//default: {cout<<"a: error word "<<input<<" processing at index "<<i<<endl;}
}
i = i + 2;
}
else if(input[i] == 'e')
{
switch(input[i+1]){
case 'e': {output =output + "E";break; }
// default: {cout<<"e: error word "<<input<<" processing at index "<<i<<endl;}
}
i = i + 2;
}
else if(input[i] == 'i')
{
switch(input[i+1]){
68
case 'i': {output =output + "I";break; }
// default: {cout<<"i: error word "<<input<<" processing at index "<<i<<endl;}
}
i = i + 2;
}
else if(input[i] == 'o')
{
switch(input[i+1]){
case 'o': {output =output + "O";break; }
// default: {cout<<"o: error word "<<input<<" processing at index "<<i<<endl;}
}
i = i + 2;
}

```

```

else if(input[i] == 'u')
{
switch(input[i+1]){
case 'u': {output =output + "U";break; }
// default: {cout<<"u: error word "<<input<<" processing at index "<<i<<endl;}
}
i = i + 2;
}
else if(input[i+1] == 'h' && len > i+ 2 && input[i+2] != 'h')
{
switch(input[i]){
case 'c': {output =output + "C";break; }
case 'd': {output =output + "D";break; }
case 'p': {output =output + "P";break; }
case 's': {output =output + "S";break; }
// default: {cout<<"h: error word "<<input<<" processing at index "<<i<<endl;}
}
i = i + 2;
}
else if(input[i+1] == 'y' && len > i+ 2 && input[i+2] != 'y')
{
switch(input[i]){
case 'n': {output =output + "N";break; }
case 'z': {output =output + "Z";break; }
// default: {cout<<"y: error word "<<input<<" processing at index "<<i<<endl;}
}
i = i + 2;
69
}
else if(input[i+1] == 's' && len > i+ 2 && input[i+2] != 's')
{
switch(input[i]){
case 't': {output =output + "T";break; }

```

```

// default: {cout<<"s: error word "<<input<<" processing at index "<<i<<endl;}
}
i = i + 2;
}
// else if(input[i+1] != "")
// {
// switch(input[i]){
// case "": {output =output + "J";break; }
// default: {cout<<"y: error word "<<input<<" processing at index "<<i<<endl;}
// }
// i = i + 2;
//}
else
{
output = output + input[i];
i++;
}
}
if (i == len -1) output = output + input[i];
return output;
}

```

```

int isVowel(char ch)
{
if(ch == 'a' || ch == 'e' || ch == 'i' ||ch == 'o' ||ch == 'u' ||ch == 'A' ||ch == 'E' ||ch == 'I' ||
ch == 'O' || ch == 'U') return 1;
return 0; }

```

Appendix D: Evaluation Questionnaire

Questionnaire

Wachemo University

Collage of Electrical and Computer Engineering

The aim of this questionnaire is to evaluate the performance of the Speech Synthesizer for Kambaata language by using statistical parametric approach. We kindly ask for you to consider each question seriously and give the grade. Please give a grade from a scale 1 to 5 in each synthesized Kambaata sentences.

Listen to the following audio files and tick (✓) in the Box what you consider is the right value for the synthesizer. Please give a grade from a scale 1 to 5 in each synthesized Kambaata sentences. 1=bad, 2=fair, 3=good, 4=very good and 5= excellent.

Part 1

1. Sex Male ☐ Female ☐

2. Age

15 – 25

26 -35

36-45

46-55

☐☐☐☐

3. Are you familiar with Kambaata Language?

Yes

☐

No

☐

4. How many years have you been a native speaker of the language

4-12

13-26

27-39

Above 40

☐☐☐☐

Part 2

How do you judge the understandability of the synthesized speech?

Sentence1. _____

Sentence2. _____

Sentence3. _____

Sentence4. _____

Sentence5. _____

How do you judge the naturalness of the synthesized speech?

Sentence1. _____

Sentence2. _____

Sentence3. _____

Sentence4. _____

Sentence5. _____

Signature

Appendix E: Sentences Used to Test the Kambaata TTS

- (uniph_001 "zakkaanchoon lameenta huje qome annannoomat ammo shumata.")
- (uniph_002 "zaba bao booru zakko fanqalanobaa jaata luushsho xawu hinaten zumanobaa.")
- (uniph_003 "zaale birriichch zaalit afuullitaa qaxat uullat woyyitaa minaaadee hasisaano xawaan.")
- (uniph_004 "yoo woma itii higgo woma saadeenno aaqqu hashshoo ilansata qooccano.")
- (uniph_005 "yoosii yoorroadaa nuurano qabaaxxaami nadassishs hoda iyyaqqeenno.")
- (uniph_006 "yooraan handa yu bareedaa hujateenan bargaqii anga aleen yoosirii maganu galaxxua.")
- (uniph_007 "xaqqumbu iibata qulxanobaa maqeechch abba ikko xawu qulxu gibu qoorimata.")
- (uniph_008 "xammi rosumbu xaaxxumbua dibata wolo keeniichch rosumbu hacalaa.")
- (uniph_009 "xalla koqanoohuu xabaru buunneeiihuu xone mexxoont hiilarra assano kenu guraan.")
- (uniph_010 "xahu xaazzeenan turka fulanobaaa bitaan waaltaa laalut abbiss qakkichchuta.")