# THE DEVELOPMENT OF ACCENTED ENGLISH SYNTHETIC VOICES

by

**PROMISE TSHEPISO MALATJI**

DISSERTATION

Submitted in fulfilment of the requirements for the degree of

**MASTER OF SCIENCE**

in

**COMPUTER SCIENCE**

in the

**FACULTY OF SCIENCE AND AGRICULTURE**

**(School of Mathematical and Computer Sciences)**

at the

**UNIVERSITY OF LIMPOPO**

**SUPERVISOR:** Mr MJD Manamela

**CO-SUPERVISOR:** Dr TI Modipa

**2019**

## DEDICATION

In memory of my grandparents, Cecilia Khumalo and Alfred Mashele, who always believed in me!

# DECLARATION

I declare that **THE DEVELOPMENT OF ACCENTED ENGLISH SYNTHETIC VOICES** is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references and that this work has not been submitted before for any other degree at any other institution.

**Signature** **Date**

# ACKNOWLEDGEMENTS

I want to recognise the following people for their individual contributions to this dissertation:

- My brother, Mr B.I. Khumalo and the whole family for the unconditional love, support and understanding.
- A distinct thank you to both my supervisors, Mr M.J.D. Manamela and Dr T.I. Modipa, for their guidance, motivation, and support.
- The Telkom Centre of Excellence for Speech Technology for providing the resources and support to make this study a success.
- My colleagues in Department of Computer Science, Messrs V.R. Baloyi and L.M. Kola, for always motivating me.
- A special thank you to Mr T.J. Sefara for taking his time to participate in the study.
- The six Computer Science undergraduate students who sacrificed their precious time to participate in data collection.
- All the volunteers who evaluated the work done in the study.
- Above all, I thank the Almighty God for gracing me with the ability to start and finish this study. And also for blessing me with the supportive people around me.

# ABSTRACT

A Text-to-speech (TTS) synthesis system is a software system that receives text as input and produces speech as output. A TTS synthesis system can be used for, amongst others, language learning, and reading out text for people living with different disabilities, i.e., physically challenged, visually impaired, etc., by native and non-native speakers of the target language. Most people relate easily to a second language spoken by a non-native speaker they share a native language with. Most online English TTS synthesis systems are usually developed using native speakers of English. This research study focuses on developing accented English synthetic voices as spoken by non-native speakers in the Limpopo province of South Africa. The Modular Architecture for Research on speech sYnthesis (MARY) TTS engine is used in developing the synthetic voices. The Hidden Markov Model (HMM) method was used to train the synthetic voices. Secondary training text corpus is used to develop the training speech corpus by recording six speakers reading the text corpus.

The quality of developed synthetic voices is measured in terms of their intelligibility, similarity and naturalness using a listening test. The results in the research study are classified based on evaluators' occupation and gender and the overall results. The subjective listening test indicates that the developed synthetic voices have a high level of acceptance in terms of similarity and intelligibility. A speech analysis software is used to compare the recorded synthesised speech and the human recordings. There is no significant difference in the voice pitch of the speakers and the synthetic voices except for one synthetic voice.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## ABBREVIATIONS AND ACRONYMS

| | | |
|---|---|---|
| CALL | – | Computer-Assisted Language Learning |
| CMU | – | Carnegie Mellon University |
| CPU | – | Central Processing Unit |
| CSIR | – | Council for Scientific and Industrial Research |
| DNN | - | Deep Neural Network |
| en_GB | – | British English locale |
| en_US | – | United States English locale |
| en_ZA | – | South African English locale |
| F0 | – | Fundamental Frequency |
| GB | – | Gigabytes |
| GHz | – | Gigahertz |
| GPU | - | Graphic Processing Unit |
| GUI | – | Graphic User Interface |
| HMM | – | Hidden Markov Model |
| HTS | – | HMM-based Synthesis System |
| HTTP | – | Hypertext Transfer Protocol |
| L1 | – | First Language |
| LTS | – | Letter-to-sound |
| MARY | – | Modular Architecture for Research on speech sYnthesis |
| MLSA | – | Mel Log Spectral Approximation |
| MOS | – | Mean Opinion Score |
| NLP | – | Natural Language Processing |
| NNS | – | Non-native speaker |

| | | |
|---|---|---|
| NS | – | Native speaker |
| NZE | – | New Zealand English |
| POS | – | Part-of-speech |
| SAE | – | South African English |
| TTS | – | Text to Speech |
| UKM | – | University of Kebangsaan Malaysia |
| US | – | United States |
| WER | – | Word Error Rate |

# CHAPTER 1

## 1. INTRODUCTION

### 1.1. Background of the Problem

Most people ideally relate and react positively to utterances by a speaker with whom they share a native (mother tongue) language. Watson *et al.* (2013) found that citizens in New Zealand responded significantly more positively to a robot if it has their New Zealand English (NZE) synthetic voice as opposed to the United States (US) accented synthetic voice from the Festival text-to-speech (TTS) online demonstration. A TTS synthesis system converts any given text in a particular language to its equivalent speech representation. A TTS synthesis system can be used to enhance second language learning or assist physically and visually challenged people.

Native speaker’s language knowledge and skills are used in developing most of the available online English TTS synthesis systems. The TTS synthesis systems are not only targeted for use by native speakers (NSs) of the target language but also intended to serve the heterogeneous group of non-native speakers (Janska *et al.*, 2010). Different English accents exist among non-native speakers (NNSs) of any target language for example, Chinese English, Indian English, and New Zealand English. There are few online English synthetic voices on Code Welt Speak online text-to-speech synthesiser that are generated from African voices, particularly from South Africans, but on the other hand, there is no online synthetic voice of different non-native accents from the South African population groups on the internet.

If online English TTS synthesis systems continue to consist of only voices from native English speakers who inevitably carry their native accents into the synthetic voices created, the growth of the TTS synthesis systems usage by non-native English speakers may be hindered. In this study, the researcher develops exotic or South African-accented synthetic voices of the English language that the majority of the country citizens will easily relate to. There is a need to develop synthetic voices targeted at South Africans as most of the official languages spoken have insufficient resources (Barnard *et al.*, 2014). Several synthetic voices for different languages (such as, isiZulu, Xitsonga, isiXhosa, Sepedi, Afrikaans and English) have been developed in the Qfrency[1] online demonstration and other TTS synthetic systems (Barnard *et al.*, 2014; Baloyi, 2012). These synthetic voices were developed using native speakers of the featured respective languages.

## 1.2. Problem Statement

As one of the most matured spoken language processing technologies, speech synthesis technology should be available in all languages existing globally. South Africa has eleven official languages (Barnard *et al.*, 2014). Many South African languages are regarded as highly under-resourced, i.e., having insufficient resources for everyday use especially in the rapid modern digital information age. Initial efforts to change this scenario are already afoot in some organisations that pursue research activities in human language technology. There are no readily available non-native accented English speech corpora and TTS synthesis systems for South African English. It is for this reason that the researcher embarked on this research project to attempt to address the problem of scarce resources in speech and language processing.

## 1.3. Aim and Objectives

The aim of this research project is to develop male and female accented English TTS synthesis systems using non-native speech data.

The objectives of this research project are:

[1] Qfrency, available on http://www.qfrency.com/demo/index.php

### 1.3.1 *Data collection*

The researcher collects and prepares the TTS training speech data from recruited volunteers using free scientific analysis of speech software, called Praat, to record the read text corpus from the Carnegie Mellon University (CMU) Arctic project.

### 1.3.2 *Voice development*

To develop and train six South African non-natives accented English synthetic voices using the Modular Architecture for Research on speech sYnthesis (MARY) TTS development engine.

### 1.3.3 *Voice testing*

To test the developed synthetic voices using secondary prompt text data from the North West University Lwazi project using a subjective listening test (evaluators are asked to rate the synthetic voices after they are subjected to synthesised speech).

### 1.3.4 *Avail voices*

Lastly, make available the developed synthetic voices for use by potential end-users such as learners from rural parts of South Africa.

## 1.4. Theoretical Framework

The Hidden Markov Model (HMM) framework, as embedded into the HMM-based synthesis system (HTS), is used to statistically model speech parameters (spectrum, phoneme duration and fundamental frequency (F0)) in training the synthetic voices developed in this research project. An HMM is a statistical time series model that is used in different field such as the speech recognition systems and the TTS synthesis systems. A Hidden Markov Model is defined as a finite state machine that generates a sequence of discrete time observations (Yamagishi, 2006). Figure 1.1 shows two examples of a typical HMM structure.

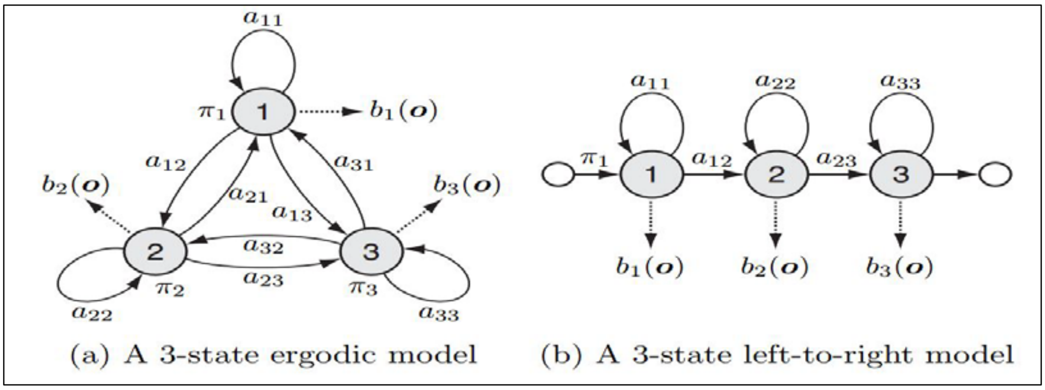


**Figure 1.1: The two examples of hidden Markov models (HMM) structure (Yamagishi, 2006).**

In Figure 1.1, (a) indicates a model in which each of the states can be reached from the other within a single transition, while the left-to-right model in Figure 1.1 (b) shows a linear model in which the state index increases or stays depending on the time increment. According to Zen *et al.*, (2007), the HMM-based synthesis system has two stages, the training stage and the synthesis stage.

### 1.4.1. *Training Stage*

With the speech database that the researcher developed, the speech analysis is performed. This phase includes a composition of context-dependent phoneme HMMs. The context-dependent phonemes are modelled using the Baum-Welch algorithm. Figure 1.2 illustrates the graphical representation of the training phase.

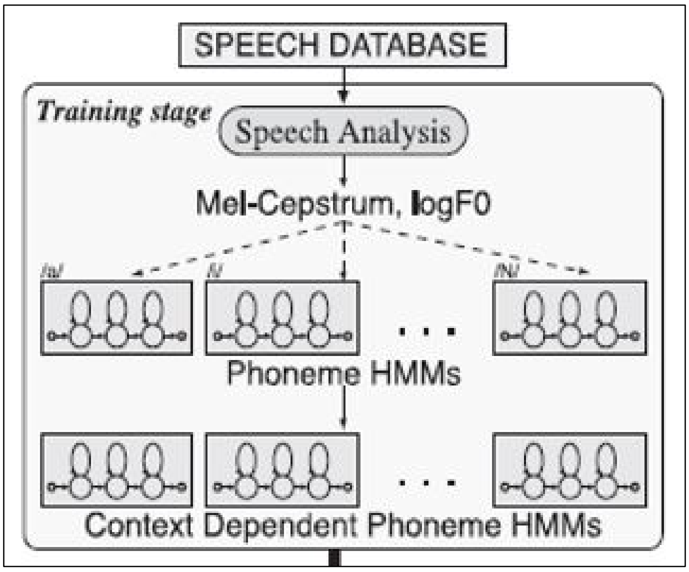


**Figure 1.2: The training stage of the HMM-based speech synthesis (Zen *at el.*, 2007).**

### 1.4.2. *Synthesis Stage*

Figure 1.3 shows how data flows from the point when the researcher input text to when speech is produced. When input text is given, it is transformed to sentence HMMs which are matched with the context-dependent phoneme HMMs from the training stage. Mel Log Spectral Approximation (MLSA) filter is used to synthesise speech from the generated mel-cepstrum and F0 parameter sequence. The synthesis stage uses both the text input to be synthesised and the context-dependent phoneme HMMs from the training stage to produce the synthesised speech.

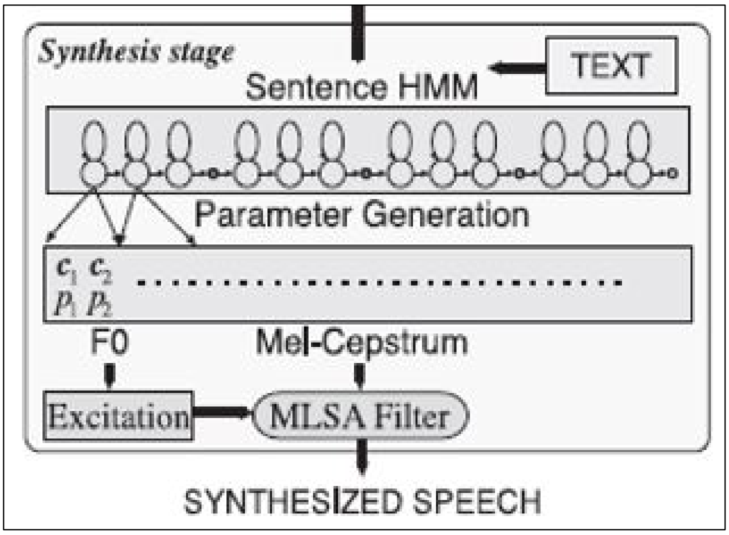


**Figure 1.3: A structural representation of the synthesis stage of HMM-based speech synthesis (Zen *et al.,* 2007).**

## 1.5. Significance of the Study

The success of the research project will have a significant contribution to the availability of South African English-accent voices for possible use in e-service delivery systems within the voice-enabled software applications. Negative perceptions that accompany South African indigenous accents when using the English language will hopefully be minimised and/or demystified. Language learning may also be eased and/or enhanced, particularly where pronunciation of proper South African names is concerned. This will further add value, not only to able-bodied end users, but also to the physically and visually challenged community of users because they will be using an English TTS synthesis system with a distinct South African accent. The research findings will likely set the trend for future researchers in the Department of Computer Science's Telkom Centre of Excellence for Speech Technology, at the University of Limpopo and other institutions of higher learning to consider further localisation of synthetic voices research to other under-resourced languages of South Africa. Since the African speakers of English in South Africa are a diverse group due to language dialects, non-native English-accented synthetic voices can further be developed for each ethnic and cultural group at any geographical location. The developed non-native English speech data and synthetic voices will be used by future researchers at the University of Limpopo and beyond.

## 1.6. Structure of the Dissertation

This dissertation is structured as follows:

In Chapter 2, the literature is reviewed or elaborated in detail. Attention is paid to current state-of-the art TTS synthesis system. The different methods that can be used to develop TTS synthesis systems are explained. The chapter takes into consideration the platform used in developing TTS synthesis systems. The usage of TTS synthesis systems by non-native speakers are discussed in detail.

Chapter 3 discusses a wide range of aspects including the chosen approach, the hidden Markov model. The development of demonstration synthetic voices, training text prompts selection and South African English locale is discussed in detail. This chapter discusses data collection and the development of the six synthetic voices and how the developed synthetic voices can be accessed from different platforms.

The evaluation of the developed TTS synthesis system is discussed in Chapter 4. Ways of recruiting potential evaluators are outlined as well as the number of evaluators used. The testing data and manner of testing is discussed and the responses from the evaluators are used to analyse the quality of the developed TTS synthesis system. The researcher gives detailed results and analysis of the experiments conducted in the study.

Chapter 5 gives the summary and concluding remarks of the entire research study. The future work for this research project is discussed.

# CHAPTER 2

## 2. LITERATURE REVIEW

### 2.1. Introduction

This chapter discusses similar work reported by other researchers. It looks at the differences between the speech by both native speakers (NSs) and non-native speakers (NNSs). Some of the recently developed TTS synthesis system are outlined, the development toolkits used, and how non-native speakers of a target language relate to TTS synthesis systems.

### 2.2. Native Speakers versus Non-native Speakers

A native speaker (NS) of a particular language is a person who has spoken that language from their childhood. The Cambridge Dictionaries Online[2] defines a NS as anyone who has talked a particular language from birth, rather than having learned the language as a child and a non-native speaker (NNS) as someone who has learned a particular language as a child or adult rather than as a baby. An NNS is not the primary speaker of that particular language, and that language is not the only language the speaker knows but can be used for communication.

The importance of the second language cannot be undermined as it involves the acquisition of a new phonological system, including new phoneme categories, phonological rules or constraints, and a new prosodic structure and the acquisition of a new social indexical system (Clopper & Bradlow, 2009). To indicate the perception of varieties of English, Clopper and Bradlow (2009) explored the awareness of the world's distinction of English using native German learners of English. The evaluators were asked to identify the dialects of English according to their country. These dialects included Northern and Southern British

[2] http://dictionary.cambridge.org/dictionary/english/native-speaker

English, Cockney English, Welsh English, Scottish English, Southern American English, Australian English, New Zealand English, South African English, West African English and Indian English. The learners were correct if they identified the right region of the world but the identification accuracy was significantly different across different varieties. The Southern American English was correctly identified by 46% of the listeners while only 3% of the evaluators could correctly identify South African English.  Clopper and Bradlow (2009) also used New Zealand evaluators who correctly characterised the New Zealand, Australian and American speakers with 85%, 57% and 66% accuracy, respectively. The Australian evaluators also successfully distinguished the New Zealand, Australian, and American talkers with 83%, 84%, and 77% accuracy respectively. These percentages indicate that it is easy for evaluators to identify the speaker they share a native language with and that native listeners are more accurate than non-native listeners.

It has been observed in studies of social and behavioural sciences that native speakers of Malay and Mandarin students or listeners have difficulties in listening to German speakers (Hassan *et al.*, 2014). Hassan *et al.* (2014) set up an experiment to determine the impact of native speech against non-native speech and 32 University of Kebangsaan Malaysia (UKM) students were used as participants. They were all taking a basic German language course of which, 16 were men and 16 were women with an average age of 22. Out of the 32 participants, 19 were of Chinese origin and thus their mother tongue is Mandarin, and the rest were Malays, with Malay language as their mother tongue. The participants were divided into two groups of 16 each; one group listened to the NNSs and the other listened to the NSs. The overall results indicated that the group that listened to non-native German speakers were found to have more score (85%) compared to those who listened to the native German speakers. From the study it was noted that students understand the NNS better than the NS.

In a multilingual country like South Africa, there are different diverse groups of NNSs of the English language and this result in different accents[3] (the way in which people in a particular area, country, or social group pronounce words). Accent variability within a particular language is a primary source of limitation of the accuracy of machine which are not adapted (Ghorshi *et al.*, 2008). According to Ghorshi *et al.* (2008), there are various factors that lead to the evolvement of accents over time, factors such as geographical variation, socio-economic classes, ethnicity, gender, age, cultural trends, mass media and immigration.

## 2.3. Text-to-speech (TTS) Synthesis System

A TTS synthesis system is defined as software that takes in the text in a particular language as input and produces its equivalent sound waveform as output (Baloyi, 2012). These synthesis systems are important not only in the physically and visually impaired community but also to those who are physically abled.

### *2.3.1. Modern TTS Synthesis Systems*

Louw, Davel and Barnard (2005) developed a general-purpose isiZulu TTS synthesis system with the aim of understanding the challenges that come with developing TTS synthesis systems for Nguni Language. They used the Festival Speech Synthesis System as the synthesis engine. To achieve state-of-the-art naturalness they used the unit selection method called Multisyn, which concatenate speech units. Multisyn required that a large text corpus be recorded. The developed isiZulu voice was found to be intelligible to most of the evaluators.

Baloyi (2012) developed a general-purpose TTS synthesis system for Xitsonga using the Hidden Markov Models (HMMs). The HMM-based speech synthesis (HTS) system produces speech that is intelligent and natural speech. In this project, a HTS toolkit was used as a patch to the HTK toolkit designed primarily

[3] http://dictionary.cambridge.org/dictionary/english/accent

for use in speech recognition (the process of receiving speech as input and produce text as output). Listeners were used to test and evaluate the developed system and most reported that the system sounded like a human being. The results showed that the TTS synthesis system built was found to be both intelligible and fairly natural.

The Council for Scientific and Industrial Research (CSIR) Meraka Institute developed a non-commercial TTS synthesis system targeted at South Africans. The online demo called Qfrency has five of the eleven official languages, namely: South African English, Afrikaans, Sepedi, isiXhosa and isiZulu. The Qfrency provides users with the opportunity to generate audio file for future use. The Qfrency TTS engine and TTS synthetic voices are available to interested users on request.

According to Walter (2016), Google's TTS tool is one of those recessive components that makes Android one of the best. This TTS system in Android phones reads contents like eBooks aloud and also enables applications to "speak" to users. Walter (2016), opines that the Google TTS comes with twenty nine different languages with no African language. The speech rate can be adjusted according to the user's preference. The inclusion of the Google TTS synthesis system in Android phones enables people who are blind to utilise their phones effectively.

### *2.3.2. Development of TTS Synthesis System*

In this section, the researcher looks at the different methods of TTS synthesis system design and their properties. The analysis of these processes will support the decision of adopting the method to be used in this project. Speech can be synthesised using two methods, the rule-driven and the corpus-based. Some methods require a set of rules that determine the synthesis: Such classifications are called rule-driven. Other methods solely depend on the recorded speech corpus that is used to synthesise the speech and concatenative synthesis and

HMM-based synthesis are examples of the corpus-driven synthesis (Huang, Acero and Hon, 2001).

- *Rule-driven Synthesis*

Articulatory and formant are typical examples of rule-driven synthesis. The articulatory synthesis is composed of three modules namely, a module for the generation of vocal tract movement, also known as the control model, a module for converting this movement information is referred to as the vocal tract model, and a module for the generation of acoustic signals is referred to as the acoustic model (Kröger & Birkholz, 2009). This method is carefully focused on the natural process of speech or singing production by people. This approach generates low-quality acoustic speech signals as compared to other acoustic speech signals generated by the corpus-based unit selection synthesis method (Kröger & Birkholz, 2009).

The formant synthesis method uses a source-filter model that fluctuates the formant frequency, amplitudes and noise to produce speech (Baloyi, 2012). It generates artificial speech waveform by fluctuating the parameters. It has the capability of producing an infinite number of speech waveforms. Its drawback though is the unnaturalness of the speech produced.

- *Corpus-driven Synthesis*

The concatenative/unit selection method was established in the 1970's and from 1980s many computer operating systems had speech synthesisers included (Sasirekha & Chandra, 2012). This method involves the use of pre-recorded human speech stored in a database. The human speech segments are concatenated to form the output speech waveform. Concatenative or unit selection involves the selection and concatenation of small units of sounds such as phonemes. A phoneme[4] is defined as a smallest structural unit that distinguishes meaning in a language whereas a phone is the instances of phonemes in the actual utterances. Phoneme are only associated with a specific language, but phones are not language-based. Concatenative synthesis has an advantage of natural sounding because it uses real recorded human voices. The drawback of concatenative synthesis is misperception from the selection of which unit to use, sentence, word or phoneme. According to Sasirekha and Chandra (2012), many TTS synthesis systems are developed using the corpus-based speech because of the high quality and natural speech output.

Black *et al.* (2007) indicate that statistical parametric synthesis has grown in popularity over the last few years. They describe it as generating the average of some set of similarly sounding speech segments. One of the cases of statistical parametric synthesis method is called HMM-based speech synthesis. For developing satisfactory speech synthesis, HMM-based synthesis system has proven to be very effective (Black *et al.*, 2007). Figure 2.1 gives the overiew of the HMM-based speech synthesis system.

[4] http://www.voxforge.org/home/docs/faq/faq/what-is-the-difference-between-a-phone-and-a-phoneme

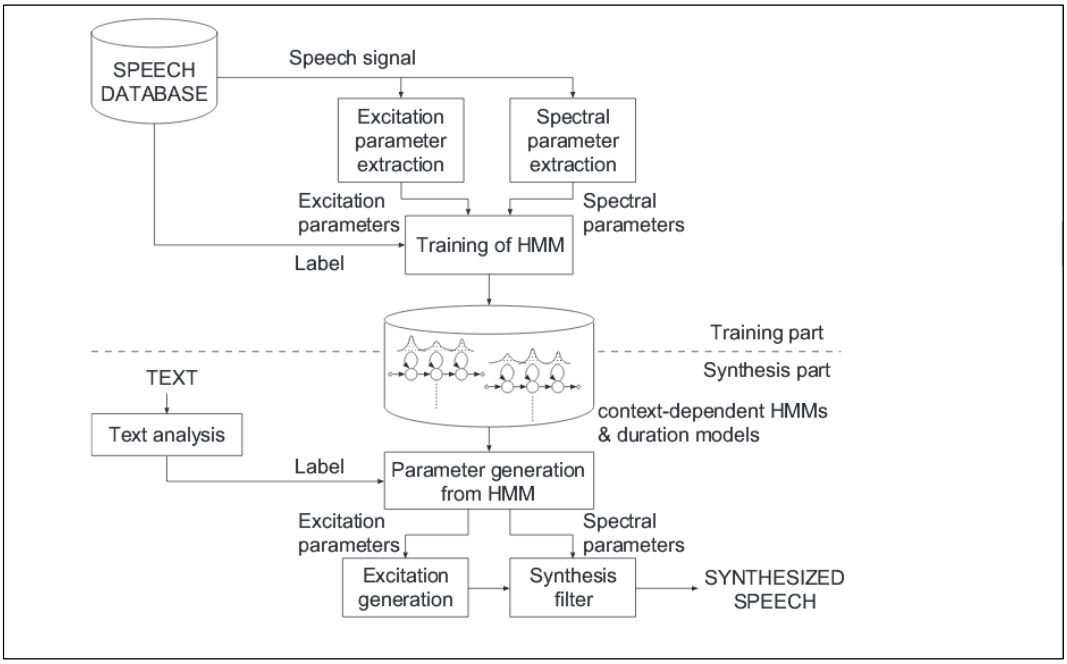


**Figure 2.1: Overview of an HMM-based speech synthesis system (Black *et al.*, 2007).**

The HMM-based synthesis system is composed of two parts, the training part and the synthesis part. Using a speech database, the speech analysis is performed. The Mel-cepstrum and fundamental frequency (F0) are extracted at each analysis frame using a continuous and multi-space probability distribution, respectively. The phoneme HMMs are modelled from the speech data and using the Baum-Welch algorithm. The re-estimation of the context-dependant phoneme HMMs is performed. The HMM and state duration models parameters are determined using the probabilistic equation:

$$\hat{\lambda} = \arg \underbrace{\max}_{\lambda} P\ (O|W, \lambda) \qquad (2.1)$$

where $\lambda$ is the model parameters, $O$ is the training data and $W$ are the transcriptions. The collection of context-dependent HMMs and state duration models are based on the maximum probability of the training data given the transcriptions and the model parameters.

When a person input text, it is transformed to context-dependent phoneme labels. Using label sequences, the sentence HMM is constructed through concatenating the context-dependent phoneme HMMs. MLSA filter is used to synthesise speech from the generated mel-cepstrum and F0 parameter sequence. The maximum probability of deciding which sequence of context-dependent HMMs to synthesis is determined by the following equation:

$$\hat{o} = arg \underbrace{max}_{o} P(o \,|w, \hat{\lambda}) \qquad (2.2)$$

where ô denotes the maximum likelihood of the synthesised speech, $o$ is the synthesised speech, $w$ is the input text and $\boldsymbol{\lambda}$ is the maximum model parameters. According to Black *et al.* (2007), there are several advantages of using the HMM-based synthesis method: The specifics of the voice can be modified easily. Its application to different languages requires less modification in the new language addition section and with small amount of speech data, one can synthesise different speaking styles or emotional speech. Its footprint is relatively small.

### *2.3.3. TTS Synthesis Platforms*

There are several TTS synthesis platforms or development engines.  The most common and widely used engines are looked at. The brief analysis of the different platforms assists the researcher in choosing the most appropriate platform to adopt.

- *Festival Speech Synthesis System*

Festival speech synthesis system supports unit selection synthesis and HMM-based synthesis. It offers a general framework for building speech synthesis systems. Festival is a free software for both commercial and non-commercial use developed in C++. The latest version of Festival includes updated HTS and CG engine, support for newer compilers and bug fixing easily. According to Pammi *et al.* (2010), Festival is the most used open source voice developing toolkit amongst the toolkits available.

- *HMM-based Speech Synthesis (HTS)*

The HMM-based speech synthesis is used to generate waveform signals from statistics of acoustic records extracted from the speech database (Morizane *et al.*, 2009). It has many attractive features as compared to the concatenative method, fully data-driven synthetic voice building, flexible synthetic voice quality control, speaker adaptation and small footprint. The HTS system is composed of two processes or phases, the training phase and the synthesis phase as illustrated in Figure 2.1. According to Morizane *et al.* (2009), the HTS system has caused dramatic improvements in the naturalness of synthetic speech compared to the alternative approach of concatenative speech synthesis. This voice developing toolkit does not include any text analyser; instead, Festival or MARY TTS system can be used with it.

- *Modular Architecture for Research on speech sYnthesis (MARY) TTS Engine*

The MARY TTS is an open source platform that is used to develop synthetic voices and is extended with the natural language processing (NLP) component (Steiner *et al.*, 2017). This platform supports the addition of new languages from scratch, like some voice developing toolkits. The MARY TTS supports both, unit selection and the HMM-based synthesis. Unlike Festival and HTS system, the MARY TTS system was developed in Java (Pammi *et al.*, 2010). According to Pammi *at el.* (2010), the MARY TTS system provides the user with graphic user interface (GUI) to lower the entrance bearer for researchers to get started with their voice developing.

- *WaveNet*

This WaveNet is the latest model of developing state-of-the-art synthetic voices (van der Oord, Dieleman & Zen, 2016). It generates raw audio waveforms through a deep neural network (DNN). This model of developing voices is fully autoregressive and probabilistic. WaveNet is known to produce more natural sounding speech than concatenative and statistical parametric (van der Oord, Dieleman & Zen, 2016). The quality of the speech produced by the WaveNet was compared to that of the concatenative, statistical parametric TTS systems and human speech based on the mean opinion score (MOS) on a 5-point scale. The results showed that WaveNet reduces the gap between the human speech and the statistical parametric and concatenative speech by over 50% for the two chosen languages (English and Mandarin).

Coto-Jiménez and Goddard-Close (2016) proposed a method of replacing the HMM with deep neural networks. The replacement brought at least one of the quality characteristics of speech synthesis, the greater naturalness and intelligibility, greater preference by users, and greater capacity to produce emotive voices (Coto-Jiménez & Goddard-Close, 2016). Le Maguer *et. al.* (2017) proposed a synchronised TTS synthesis system to compare the two standard methodologies, HMMs and DNNs. In the evaluation the DNNs outperformed the HMMs even though less than two hour of data was used. According to Qian and Soong (2014) the DNN training improved significantly since 2006 by using the computationally powerful graphics processing unit (GPU).

- *Other Open Source Voice Developing Toolkits*

The voice developing toolkits mentioned here are discussed in detail in (Pammi *et al.*, 2010).

- MBROLA system is a speech synthesiser based on concatenation of diphones. For new voice developing, diphone database must be provided to the MBROLA owners which will process and adapt it to the system format at no costs.

- FreeTTS is written entirely in Java language and is based on the CMU Flite engine (the lite version of Festival and Festvox).
- Epos is a rule-driven TTS system designed specifically for research purpose. It supports Czech and Slovak languages, and it also can be used as a front-end for the MBROLA diphone synthesiser.
- eSpeak is a formant synthesiser developed in C++ to support several languages. Like Epos, eSpeak can serve as a front-end for MBROLA.
- Gnuspeech is an articulatory TTS system that includes a GUI-based database developing. It compiles for Mac OS/X and GNU/Linux under GNUStep.

Based on the analysis of different methods of developing synthetic voices overviewed, the usage of DNNs in WaveNet generates state-of-the-art synthetic voices. The need of a GPU computer made it difficult for to use the DNNs in training the synthetic voices.

### *2.3.4. Usage of TTS Synthesis System*

The TTS synthesis systems are used in different situations for different reasons and with the rapid increase in the quality of the synthetic voices, the application fields also increases steadily. The different applications fields[5] of the TTS synthesis technology include: applications for the blind, deafened and vocally handicapped, education, telecommunications and multimedia. A TTS synthesis system has the capability[6] to be used also in business, academic, government and disability applications.

[5] http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap6.html

[6] http://savoices.inclusivesolutions.co.za/how-does-it-work/

### *2.3.5. TTS Synthesis System Evaluation*

There are several methods of evaluating synthetic voices, the objective[7] (not influenced by personal opinion or believe but based on real facts) evaluation and the subjective[8] (based on personal opinion or believe than real facts) evaluation. The most common method of evaluating TTS synthesis system is the subjective evaluation through the listening test (Lemmetty, 1999). In the study by Thomas (2007) the listening test was conducted to evaluate the developed TTS synthesis system using 20 evaluators. To evaluate the Xitsonga TTS synthesis system, 16 evaluators were recruited with both genders evenly represented (Baloyi, 2012).

### *2.3.6. TTS Synthesis System for Non-native Speakers*

According to Janska *et al.* (2010), the TTS synthesis systems are often provided in English for the heterogeneous targeted group of users even though their native language (first language also known as L1) is not English. The NNSs of a language (English in this study) are a diverse group, and there is no single English TTS synthesis system that can be well accepted by the whole NNSs group. Developing TTS synthesis systems using NNSs of that language is likely to improve the acceptability of applications of the TTS synthesis systems to the extent that potential users will listen to the synthesised voice language with an accent they relate to (Oshima *et al.*, 2015). As a result of the communication break down between native Japanese speaker and a native English speaker due to Japanese-accented prosody, the demand in Japan for Computer-Assisted Language Learning (CALL) is targeted at bridging the break down.

The importance of accent is indicated by the need of New Zealand-accented voices because blind New Zealanders are listening to foreign accents and that causes them to lose their identity (McAvinue, 2014). Most of the New Zealand TTS synthesis systems came with the American accent as a default standard. A

[7] http://dictionary.cambridge.org/dictionary/english/objective
[8] http://dictionary.cambridge.org/dictionary/english/subjective

New Zealand synthetic voice to be developed is believed to have the potential to protect the New Zealand identity (McAvinue, 2014). The co-director of the Blind Sight was exposed to different synthetic voices in English (accents from Wales, Ireland, India and South Africa) but none of the synthetic voices was in his accent. English is one of the eleven official languages of South Africa pre-1994 and post-1994, one accent to produce synthetic English voices does not significantly accommodate the majority group of non-native speakers of English. The uniqueness[9] of having eleven official languages inspires the need to develop South African synthetic voices, locally and internationally.

## 2.4. Summary

The researcher presented speech by native and non-native speakers and how each perceives the speech by the other. The modern TTS synthesis systems were discussed. The methods of developing synthetic voices were outlined as well.

[9] Qfency: http://www.qfrency.com/

# CHAPTER 3

# 3. THE MARY TTS SYNTHESIS SYSTEM AND VOICE DEVELOPMENT EXPERIMENTS.

## 3.1. Introduction

For one to develop synthetic voices using the DNN model, one requires a more computationally powerful machine than the standard desktop computer architecture used in this research project. The Festival, HTS and MARY TTS synthesis systems can be adopted and used on the standard desktop computer architecture. As a result, the statistical parametric synthesis is used in the development of our synthetic voices. The MARY TTS synthesis engine platform is used to develop and train the synthetic voices.

In this chapter the researcher presents the preparation of the MARY TTS synthesis engine to be used in developing the targeted accented synthetic voices. All the main and extra software packages used in the installation are given. The demonstration synthetic voices are developed to test the functionality of the engine and to assist the researcher in choosing the training text corpus. The researcher presents the process of developing a new South African English locale to be used in the development of the accented synthetic voices. The data collection and synthetic voice development processes for each speaker is fully discussed. The steps of accessing the new synthetic voices through a Microsoft Windows operating system are outlined.

## 3.2. MARY TTS Synthesis System and Software Packages

This section explains the development of a MARY TTS on a desktop workstation. The operating system used is Ubuntu 14.04 LTS 32-bit on a desktop with 2 Gigabytes (GB) memory. The computer operates on an Intel® Core™2 Duo CPU

E7500 @ 2.93 GHz x 2 processor with a storage capacity of 500 GB. The packages in Section 3.2.1 were needed before the installation process. All the downloaded packages must be placed in the */voice/source* directory and for installing MARY TTS synthesis system we followed the steps on New Voice Creation tutorial (Laura, 2015). *The software packages needed to install MARY TTS synthesis system are:*

**Apache-Maven** – is a software project management and comprehension tool, used to manage project's build and documentation.

**Audacity** – is a free, open source, cross-platform audio software for multi-track recording and editing.

**cmu_us_slt_arctic** – TTS speech data that contains 1132 utterances spoken by a US English female speaker. The speaker used is experienced in developing the synthetic voice.

**Festival** – is a standard multilingual system that affords a platform for developing TTS systems. There are many voices and lexicons available for download that can be used with festival.

**Festvox** – repository which is aimed at making the task of building a new voice easy by providing example speech databases.

**HDecode** – is decoder used by HTK to handle large vocabulary using cross-word triphone models[10]. One needs to register as an HTK user, and also agree to its licence to download it.

**HTK** – is a toolkit that was primarily designed for use in speech recognition research for building and manipulating hidden Markov but can now be used in many different applications including speech synthesis.

**HTS** – is an HMM-based speech synthesis toolkit with a training part developed as a modified version of the Hidden Markov Model Toolkit (HTK) and it cannot be used alone. HTS must be patched to HTK and the license terms and condition of HTK must be adhered to after patching.

[10] http://www.seas.ucla.edu/spapl/weichu/htkbook/node52_mn.html

**hts_engine** – is software that is used to synthesise speech waveform from HMMs output by HTS.

**Praat** – is a free software for sound operation, phonetic analysis and acoustic analysis and reconstruction of speech signals.

**Speech tools** – is a collection of C++ functions for the speech processing of related speech objects, and is used for reading, writing, converting and supporting speech processing objects such as fundamental frequency, waveform, labels, etc..

**SPTK** – is a software package that comprises speech signal processing tools.

**tts.lwazi.eng –** includes 447 utterances spoken by a native SA English male speaker.

## 3.3. Synthetic Voices Experiments

After the installation we did not have any audio files to test the system. Hence we used the Arctic data from Carnegie Mellon University (CMU). We developed four demonstration voices to test the installed system and help us make informed decisions on the type and amount of text prompts to use.

### *3.3.1. voice_slt*

The training speech data for this voice was downloaded from CMU website[11] and we unpacked it in the */voice/data/* directory. These recordings were produced by a female native speaker of English, Stefanie L. Tomko (slt) from the United States (US). This text corpora consists of 1132 prompt sentences, covering 10175 words out of which 2974 are unique words. The total number of phones which are covered in this text corpus is 39153 (Kominek *et al.*, 2004). We then developed

[11] http://www.speech.cs.cmu.edu/cmu_arctic/packed/cmu_us_slt_arctic-0.95-release.tar.bz2

the English synthetic voice using this readily available data. The same synthetic voice can be accessed through the marytts-client.

To install the first synthetic voice the researcher acquired and executed the script and made changes to *voice-slt-hsmm-5.2-SNAPSHOT-component.xml*:

gender = “female”

name = “*voice_slt*”

description: A female English general voice

After the changes the researcher acquired and executed the *installer.sh* script to install the new synthetic voice. Once the voice is successfully installed onto the MARY TTS, it could be accessed through a web browser on the localhost and the port 59125, i.e., localhost: 59125 and the MaryTTS Web Server reflect the voice installed as a default voice. To vary our voice options and test the functionality of the system to incorporate multiple voices, three more voices were developed.

### 3.3.2. voice_lwazi

The researcher used an English TTS corpus obtained from Lwazi project to develop the second voice named *voice_lwazi*. The audio files (utterances) spoken by a South African male NS of English reading out 447 sentences (3834 words). Another directory was developed inside the *voice/* directory and named it data1. This directory contained two sub-directories, the wave and the text directory. The audio files were place in *voice/data1/wav* and the *txt.done.data* in the directory *voice/data1*. The voice was then followed with some changes to the directory paths and in some properties relating to the speaker. The auto labelling and training of the second voice did not take more time like the synthetic voice *voice_slt* did; this is because the first voice had more training data compared to the second voice. To install the second synthetic voice, the researcher executed the *copying.sh* script and made changes to *voice-lwazi-hsmm-5.2-SNAPSHOT-component.xml*:

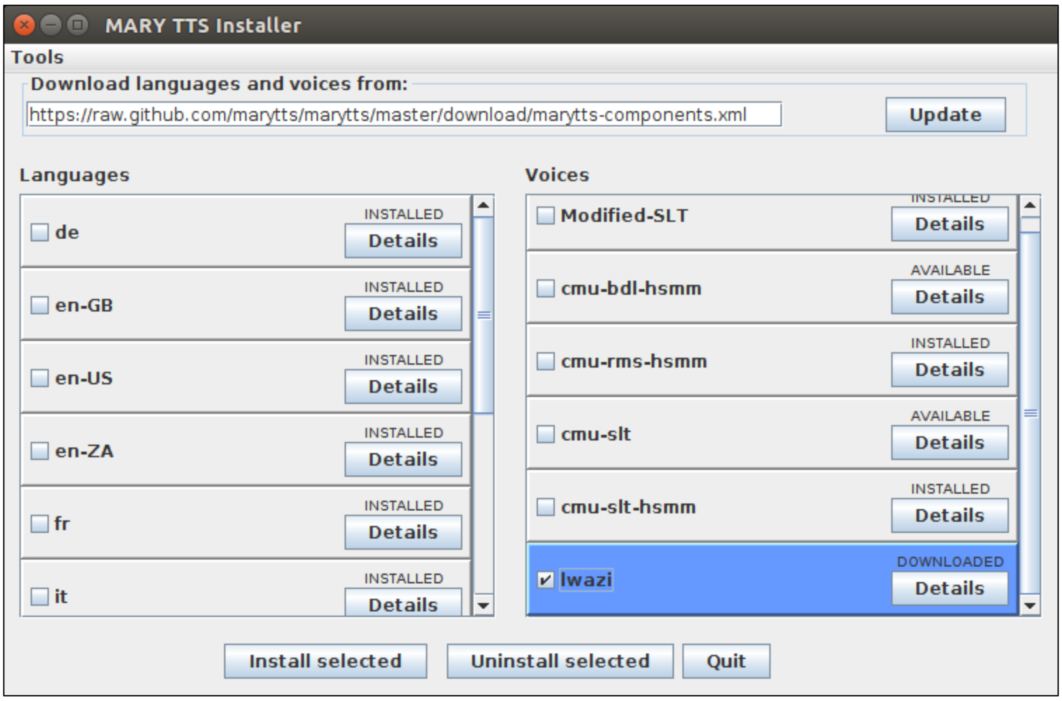
gender = “male”

name = “*voice_lwazi*”

description: A male English general voice

After the changes the *installer.sh* script was executed to install the new synthetic voice voice_lwazi.

Figure 3.1 shows the screen dump to install and remove components. The new *voice_lwazi* voice was selected for installation onto the MARY TTS.



Download languages and voices from: https://raw.github.com/marytts/marytts/master/download/marytts-components.xml

Languages

- □ de INSTALLED Details
- □ en-GB INSTALLED Details
- □ en-US INSTALLED Details
- □ en-ZA INSTALLED Details
- □ fr INSTALLED Details
- □ it INSTALLED Details

Voices

- □ Modified-SLT INSTALLED Details
- □ cmu-bdl-hsmm AVAILABLE Details
- □ cmu-rms-hsmm INSTALLED Details
- □ cmu-slt AVAILABLE Details
- □ cmu-slt-hsmm INSTALLED Details
- ☑ lwazi DOWNLOADED Details

**Figure 3.1: The snapshot of MARY TTS installer for installing synthetic voices and languages.**

The two new synthetic voices were used to synthesise speech and it was noted that the *voice_slt* synthetic voice sounds more natural than the *voice_lwazi* synthetic voice. The cause of the difference in the quality of the synthetic voices was unknown. The impact of the amount of the training speech data on the quality of the synthetic voice was investigated through the development of another two new synthetic voices, *voice_rms* and *voice_sltmodified*.

### *3.3.3. voice_rms*

The third synthetic voice named *voice_rms* was developed using the training speech data obtained from the CMU. Richard M. Stern (rms) is a male native speaker of English from the US. The data used to record the wave files is the same text data as the *voice_slt* one, with 1132 sentences. The new synthetic voice nemed *voice_rms* produced intelligible and natural speech.

### *3.3.4. voice_sltmodified*

The fourth voice named *voice_sltmodified* was developed by modifying the *voice_slt* training data. We modified the training speech data of *voice_slt* by selecting some audio files and their transcriptions. We took the first 447 wave files with their corresponding sentences as our training data. Inside the *voice/* directory, we created the *data2* directory which also had its subdirectories, *voice/data2/wav/* and *voice/data2/*. The new synthetic voice was trained following the voice prepare, auto labelling and training as shown in Appendix A with some changes to the directory paths and properties of the speakers. The synthetic voice *voice_sltmodified* was found to be not as natural as the synthetic voices *voice_slt* and *voice_rms*. After few volunteers listened to the four synthetic voices, it was noted that voices trained with 1132 recorded sentences have better quality than those trained with 447. Therefore, the CMU arctic data was adopted for our recording to achieve better quality synthetic voices.

## 3.4. South African English

The MARY TTS system has two types of English locales as default, the United States English (en_US) and the British English (en_GB). Neither one of the English languages already available has all the phones spoken in South African English (SAE). Hence we created a new locale for the South African English (en_ZA) as follows:

Using the Lwazi phone set, the *allophones.en_ZA.xml* file in appendix D was created manually and table 3.1 has all the SAMPA phones that are only in SAE.

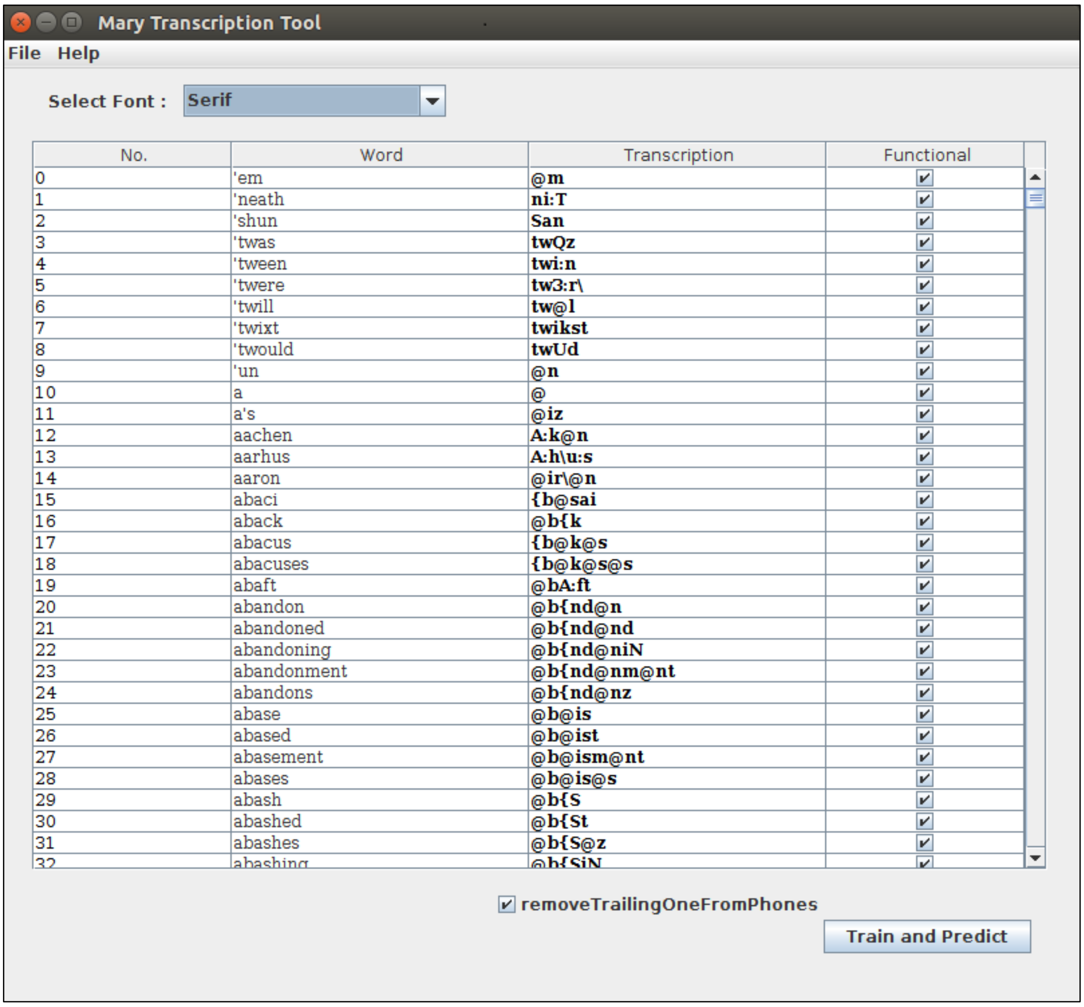
**Table 3.1: List of additional phones applicable to South African English.**

| Type | SAMPA | IPA | SAE phones |
|---|---|---|---|
| Vowels | i: u: 3: O: a A: Q | iː uː ɜː ɔː a ɑː ɒ | A O u i i: u: { V E I U 3: O: a A: Q @ r= aU OI @U EI AI |
| Affricates consonants | d_0Z | dʒ | tS dZ d_0Z |
| Fricatives consonants | x h\ | x ɦ | f v T D s z S Z x h h\" |
| Diphthongs consonants | @i ai Oi @u au i@ e@ u@ | əi ai ɔi əu au iə eə uə | @i ai Oi @u au i@ e@ u@ |
| Approximant consonants | r\ l | ɹ l | r r\ w j l |
| Nasal consonants | | | m n N |
| Stop consonants | | | p t k b d g |

Before we proceeded with the new language creation, mysql was installed using the following command:

*sudo apt-get install mysql*

The transcription tool is used for transcribing new language text corpus (*en_ZA.txt*) and automatic training of letter-to-sound (LTS) rules for the SAE to be used in tokenisation. The transcription tool uses all functional words in the SAE to build a primitive part-of-speech (POS) tagger. It also develops a pronunciation dictionary. The Figure 3.2 depicts the screen dump of the transcription tool interface with the SAE functional words and their transcriptions.



Select Font : Serif

| No. | Word | Transcription | Functional |
|---|---|---|---|
| 0 | 'em | @m | ☑ |
| 1 | 'neath | ni:T | ☑ |
| 2 | 'shun | San | ☑ |
| 3 | 'twas | twQz | ☑ |
| 4 | 'tween | twi:n | ☑ |
| 5 | 'twere | tw3:r\ | ☑ |
| 6 | 'twill | tw@l | ☑ |
| 7 | 'twixt | twikst | ☑ |
| 8 | 'twould | twUd | ☑ |
| 9 | 'un | @n | ☑ |
| 10 | a | @ | ☑ |
| 11 | a's | @iz | ☑ |
| 12 | aachen | A:k@n | ☑ |
| 13 | aarhus | A:h\u:s | ☑ |
| 14 | aaron | @ir\@n | ☑ |
| 15 | abaci | {b@sai | ☑ |
| 16 | aback | @b{k | ☑ |
| 17 | abacus | {b@k@s | ☑ |
| 18 | abacuses | {b@k@s@s | ☑ |
| 19 | abaft | @bA:ft | ☑ |
| 20 | abandon | @b{nd@n | ☑ |
| 21 | abandoned | @b{nd@nd | ☑ |
| 22 | abandoning | @b{nd@niN | ☑ |
| 23 | abandonment | @b{nd@nm@nt | ☑ |
| 24 | abandons | @b{nd@nz | ☑ |
| 25 | abase | @b@is | ☑ |
| 26 | abased | @b@ist | ☑ |
| 27 | abasement | @b@ism@nt | ☑ |
| 28 | abases | @b@is@s | ☑ |
| 29 | abash | @b{S | ☑ |
| 30 | abashed | @b{St | ☑ |
| 31 | abashes | @b{S@z | ☑ |
| 32 | abashing | @b{SiN | ☑ |

☑ removeTrailingOneFromPhones

**Figure 3.2: The snapshot of MARY transcription tool for transcribing text corpus for a new language.**

The phoneset file (*allophones.en_ZA.xml*) created is loaded first. The *en_ZA.txt* file is opened and all the functional words are selected on the GUI. The United States and British English voices are already existing in MARY TTS synthesis system, as such one of the projects already available (British English project (gb)) is used as a reference. The *en_ZA.config*, *allophones.en.ZA.xml* and *en_ZA.txt* files are opened and edited using the *languagefiles.sh* script.

The transcription tool will develop the following files in *voice/source/marytts/marytts-languages/marytts-lang-en/lib/modules/en/za/lexicon*:

- *en_ZA.lts*
- *en_ZA_lexicon.dict*
- *en_ZA_lexicon.fst*
- *en_ZA_pos.fst*
- *en_ZA_pos.list*

The en_GB directories and the files were copied and edited, renaming GB with ZA. A default text for the SAE was created. The en_ZA default text is set to be "Welcome to South African English!" and changes were also made in the *TOKENS_en_ZA.example* file to accommodate the new default text. The file in */en_config* is edited by effecting the changes in Appendix B. The *languageinstallation.sh* script was executed to test and install the new language file.

The *marytts-lang-en-SNAPSHOT.jar* file in the directory */marytts-lang-en/target* was copied to */marytts-5.2.SNAPSHOT/download/*. The marytts-component-installer was opened using *installer.sh*. The *marytts-component-installer* opens a graphic user interface (GUI) with en_ZA appearing as downloaded. After installing, the marytts server is restarted using *server.sh*, and the new language will be incorporated.

When users launch the Mary Web Client, to check the available locales, open the “interactive documentation of the HTTP interface to MARYTTS” hyperlink and select the locales link.

### 3.5. Data Collection

The majority of South Africans are non-native speakers of English. A total of 1132 sentences from the CMU Arctic data were acquired to create a training speech data set. The recruited speakers used to collect data are native speakers of three different South African indigenous languages, namely, Xitsonga, Tshivenda and Northern Sotho. The speakers are undergraduate students from 18 years to 25 years. They must have attended a public school and never taught English by a native speaker of English. The Xitsonga language is spoken in a wide area of the South-Eastern part of the Southern Africa (Zerbian, 2007). The Tshivenda language is spoken by many people from the northern Transvaal South Africa. The Tshivenda language speaking people occupied the south land of Limpopo between the 17th century and early 18th century (Madiba, 1994). Northern Sotho is one of the nine indigenous languages of South Africa, spoken mostly by people living in the northern Transvaal of the country (National African Language Resource Center, 2015). All the three indigenous languages are dominant in the northern part of Transvaal in South Africa, currently known as the Limpopo province. Figure 3.3 shows the percentages of native speakers of all the eleven official languages in South Africa. The Sepedi language, Xitsonga language and Tshivenda language are the fifth, eighth and tenth most spoken languages in South Africa, respectively.

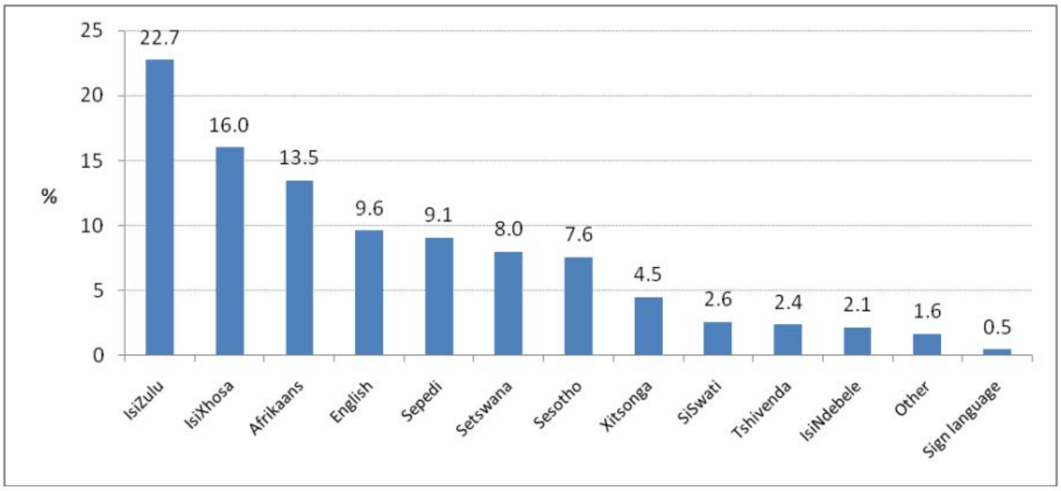


**Figure 3.3: The percentage of the first language speakers of South African language speakers, (Lehohla, 2012).**

Figure 3.4 shows the percentages of first language speakers in Limpopo provinces. The most spoken languages in Limpopo province are namely, Sepedi with 52.9%, Xitsonga with 17% and Tshivenda with 16.7%. These three languages (Sepedi, Xitsonga and Tshivenda) are mostly spoken in Limpopo province than all other eight provinces, (Lehohla, 2012).

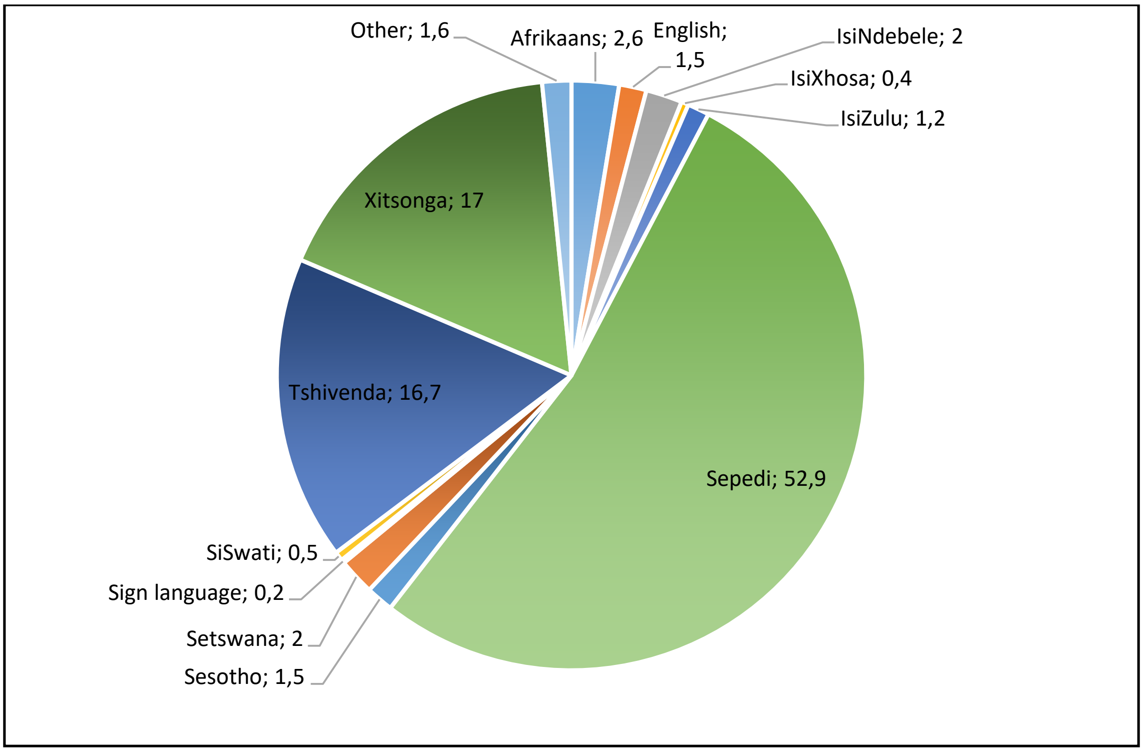
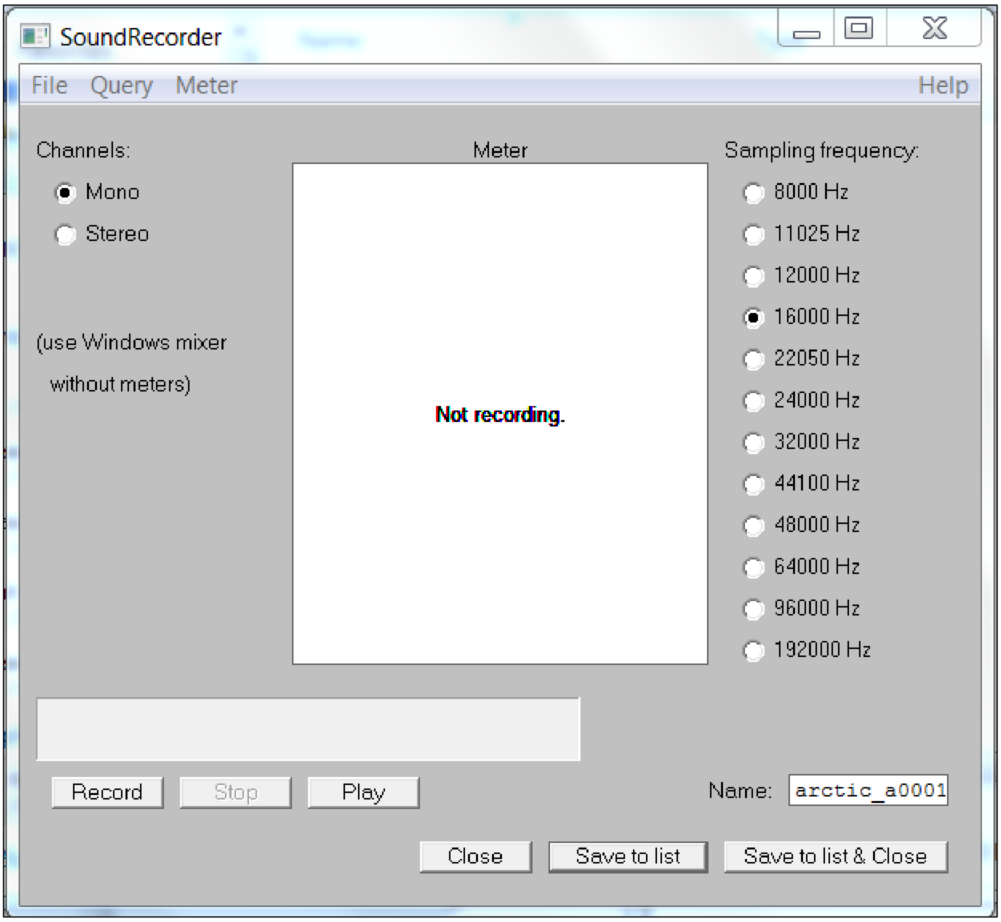


**Figure 3.4: Population percentage by first language in Limpopo province, (Lehohla, 2012).**

The read speech recordings were done at different times but the same environment. All recordings were done in a relatively quiet office. The speakers used for this project have no experience in the building synthetic voices. The Praat speech analysis tool was used to record our sentences. The channel was set to mono, and the sampling frequency was 16000 Hz for all speakers. For naming the objects or wave files, we used the Arctic naming format, i.e. *arctic_a000** and *arctic_b000** as the training text corpus is from the CMU Arctic project. A desktop microphone was used to collect the training speech data. Figure 3.5 depicts a screen dump of the Praat speech recording software and the parameters used. The personal and recording characteristics of the respective speakers are given in detail this section.



Channels:
● Mono
○ Stereo

Sampling frequency:
○ 8000 Hz
○ 11025 Hz
○ 12000 Hz
● 16000 Hz
○ 22050 Hz
○ 24000 Hz
○ 32000 Hz
○ 44100 Hz
○ 48000 Hz
○ 64000 Hz
○ 96000 Hz
○ 192000 Hz

Not recording.

Name: arctic_a0001

**Figure 3.5: Praat sound recorder interface screen dump.**

The read speech corpus procedure was followed in developing the speech database. All our speakers were recorded reading the same text corpus. Table 3.2 shows the number of sessions per speaker and the number of sentences and words each speaker read per session.

### *3.5.1. Xitsonga Female*

The speaker was 23 years old. She is born in a unilingual rural settlement called Gandlanani in Malamulele, in the Vhembe district, of Limpopo province. She attended both her primary and secondary education in public schools and was never taught English or any subject by an English native speaker. The speaker had to read out more louder to produce more audible recordings. She was allowed to listen to the last five recordings from the previous session to adjust her voice accordingly. Her reading was moderate although some words were unfamiliar to her and she committed some reading errors.

### *3.5.2. Tshivenda Female*

Due to the unavailability of volunteers, the volunteer we managed to secure her service is a year younger than the minimum age of 18 years. This young lady is from Duthuni in Thohoyandou, also in the Vhembe district, Limpopo. Duthuni is a rural area in the Northern Province of South Africa. Like the first speaker, she attended both her primary and secondary education in public schools and was never taught English by a native speaker. Her reading skill is good, with a fast reading pace than all other speakers. Like any other speaker, some words were unfamiliar to her, but she committed few errors. The speech rate of this speaker is very high and she has the minimal recording time per sentence. This speaker and the other female speakers had few errors, unlike the male speakers.

### *3.5.3. Sepedi Female*

Due to the intensity of the recording, as some recruited volunteers wanted to be paid for their services, and that resulted in many speakers in this category withdrawing. As a result, the services of a semi-rural (township) Northern Sotho female speaker were acquired. The speaker originates from the Capricorn district of Limpopo Province, in Mankweng township. She was never taught English by a native speaker during her schooling in a private primary school and public secondary school. She is fluent in speaking and reading English with fewer mistakes than other speakers. She might not have a comparable reading speed as the Tshivenda female, but her reading speech is normal.

### *3.5.4. Xitsonga Male*

The speaker is 25 years old and from a community in Mpumalanga, just next to the provincial demarcation line between Limpopo and Mpumalanga province. He is from the Bohlabela district, in a rural settlement called Acornhoek (Bushbuckridge). He attended his basic education in public primary and secondary school. He is currently in his first level, pursuing a BSc in Mathematical Sciences. His reading skill is moderate, with average reading errors.

### *3.5.5. Tshivenda Male*

The speaker is 20 years old in his first level of study pursuing a BSc in Mathematical Sciences. He is from a Tshivenda rural area called Tshipise in Thohoyandou, Vhembe district, Limpopo. He attended both primary and secondary education in public schools and for his entire education he was taught English by a non-native speaker. He has a moderate reading skill with more errors committed. His reading speed is below moderate. The speaker had to rerecord the same text prompts more than all the other speakers, and the speech data was obtained after four trials of recording. Due to the reading errors observed, and we had to discard the previous recordings and recruit a proof-reader for this speaker to minimise these errors. Table 3.3 indicates that this speaker had the longest recording time.

### *3.5.6. Sepedi Male*

The speaker from Moletjie (Leokama) in the Capricorn district of Limpopo is 22 years of age. Moletjie is one of the rural areas in the Limpopo province. He is currently in his first level of tertiary study, pursuing a BSc Mathematical Sciences. He acquired his education, both primary and secondary from public schools in Moletjie village.

**Table 3.2: The recording details for all sessions per speaker indicating the time taken per session, number sentences words covered.**

| | Session | Length (minutes) | Sentences | Words |
|---|---|---|---|---|
| **Xitsonga female** | First | 210 | 359 | 3254 |
| | Second | 125 | 234 | 2109 |
| | Third | 120 | 267 | 2389 |
| | Fourth | 135 | 272 | 2423 |
| | | | | |
| **Tshivenda female** | First | 150 | 501 | 4526 |
| | Second | 130 | 431 | 3832 |
| | Third | 55 | 200 | 1817 |
| | | | | |
| **Sepedi female** | First | 140 | 400 | 3633 |
| | Second | 120 | 317 | 2872 |
| | Third | 120 | 415 | 3670 |
| | | | | |
| **Xitsonga male** | First | 90 | 215 | 1961 |
| | Second | 90 | 144 | 1293 |
| | Third | 255 | 501 | 4498 |
| | Fourth | 160 | 272 | 2423 |
| | | | | |
| **Tshivenda male** | First | 240 | 359 | 3254 |
| | Second | 180 | 234 | 2109 |

|  |  |  |  |  |
|---|---|---|---|---|
|  | Third | 120 | 150 | 1381 |
|  | Fourth | 180 | 200 | 1721 |
|  | Fifth | 150 | 189 | 1710 |
|  |  |  |  |  |
| **Sepedi male** | First | 120 | 142 | 1308 |
|  | Second | 240 | 575 | 5197 |
|  | Third | 155 | 415 | 3670 |

Most of the speakers had a tendency of resuming all the recordings at a high note and it will gradually decrease. The map in Figure 3.6 indicates the respective geographical locations with red marks where the respective speakers originate from within Limpopo province:

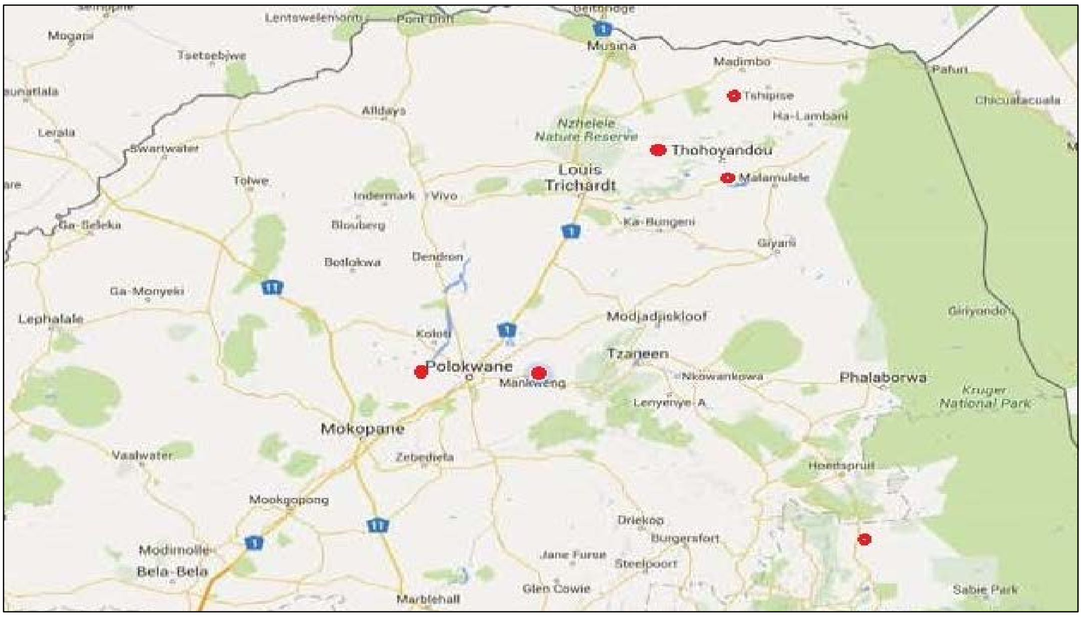


**Figure 3.6: A Limpopo province map with the six speakers' places of origin marked.**

## 3.6. Development of Accented Voices

The voices developed are named after the respective speakers' native language and their gender as it appears in Table 3.4. The number of speech training data used is the same for all the synthetic voices except for the Tshivenda female synthetic voice. A recording from the training speech data of the Tshivenda female had a mispronunciation, as such, it had to be discarded. Other voices were developed using original recordings, and only one synthetic voice was developed with the training speech data refined. The original speech training data for Tshivenda female had too much noise and the training stage failed.

**Table 3.3: The total duration per speaker, number of sentences used and the synthetic voice naming.**

| Speaker | Recording time (hours) | Sentences recorded | Sentences used | Voice name |
|---|---|---|---|---|
| Xitsonga female | 9.83 | 1132 | 1132 | xitsonga_female |
| Tshivenda female | 5.58 | 1132 | 1131 | tshivenda_female |
| Sepedi female | 6.33 | 1132 | 1132 | sepedi_female |
| Xitsonga male | 9.92 | 1132 | 1132 | xitsonga_male |
| Tshivenda male | 14.50 | 1132 | 1132 | tshivenda_male |
| Sepedi male | 8.58 | 1132 | 1132 | sepedi_male |

### *3.6.1. Voice xitsonga_female*

The synthetic *xitsonga_female* voice was developed using the original speech data obtained from Xitsonga female speaker and it was the first synthetic voice to be developed. Unlike the demonstration synthetic voices developed in the study, this voice and all other voices are trained using the en_ZA language. A directory named *xitsonga_female* was created inside the voice directory. The properties in the Database import for this synthetic voice are amended with the Duration Threshold set to 10.

HMMVoiceConfigure Settings Editor set according to the properties for female speaker as recommended by the MARY TTS:

HMMVoiceConfigure.LowerF0 = 80

HMMVoiceConfigure.mgcBandWidth = 24 (for cepstral form)

HMMVoiceConfigure.mgcOrder = 24 (for cepstral form)

HMMVoiceConfigure.UpperF0 = 350

The above settings are applicable to all the female voices.

HMMVoiceConfigure.speaker = *xitsonga_female*

After copying all the *SNAPSHOT-component.xml* files using the *copying.sh* script, the following language setting is applicable to all the accented synthetic voices to:

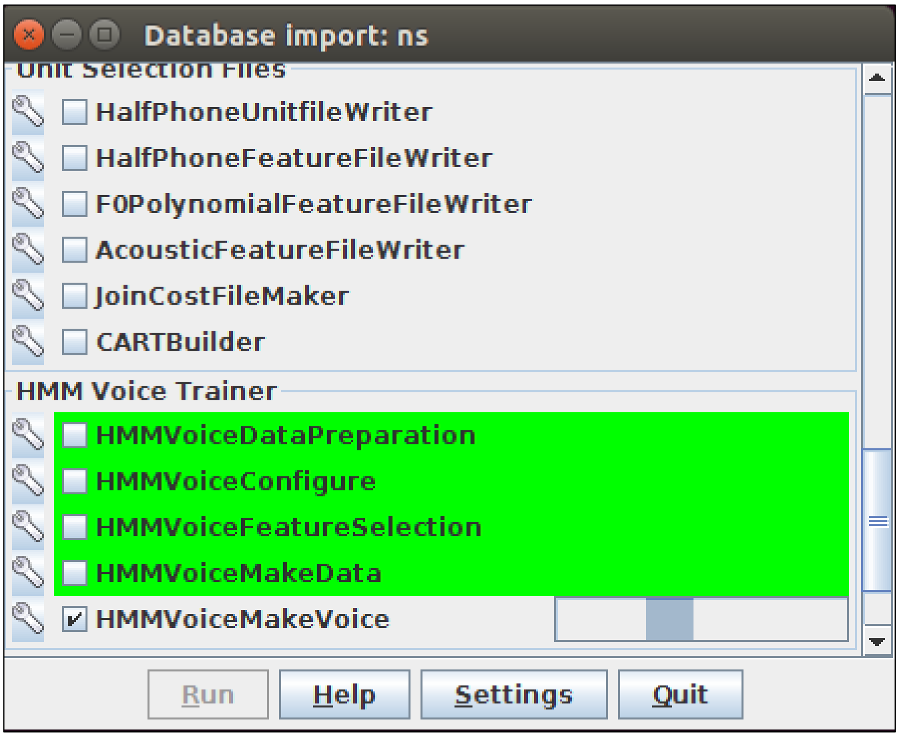
locale = “en_ZA”

language = “en-ZA”

Then the voice name and description for our synthetic voice will be:

name = “*xitsonga_female*”

description: “A Xitsonga speaking female English general voice”



Unit Selection Files

□ HalfPhoneUnitfileWriter
□ HalfPhoneFeatureFileWriter
□ F0PolynomialFeatureFileWriter
□ AcousticFeatureFileWriter
□ JoinCostFileMaker
□ CARTBuilder

HMM Voice Trainer

□ HMMVoiceDataPreparation
□ HMMVoiceConfigure
□ HMMVoiceFeatureSelection
□ HMMVoiceMakeData
☑ HMMVoiceMakeVoice

**Figure 3.7: The snapshot of HMM Voice Trainer phase of the synthetic voice development procedure.**

Figure 3.7 shows the HMM Voice Trainer stage of the synthetic voice development procedure. The training procedure takes several hours, for this synthetic voice it took approximately six hours (from 08:31 to 14:15). The time required to develop all the synthetic voices has been found to be approximately equal, including the preparation and auto labelling, the whole synthetic voice development process was observed to be about nine hours. During the training phase, which took roughly six hours, insertion, deletion or substitution errors are detected in the speech training data. Insertion error – an error caused by the speaker saying a word when it was not spoken. Deletion error – an error caused by the speaker omitting a word from the sentence(s) read. Substitution error – an error caused by the speaker replacing a word in the sentence(s) with his/her word. A third person was recruited to proofread the text training data while the speakers are recording. This process helped in eliminating errors as it was done per sentence for every speaker. If there is any mismatch between the audio and transcription, it will be detected during the stage of alignment and making global variance and the training phase will not complete successfully. It is only when there is no error that synthetic voice will train to completion. The voice or component installation is followed as in the *lwazi* voice to install this voice to the MARY Web Client.

### *3.6.2. Voice tshivenda_female*

This is the only synthetic voice that was developed using refined training speech data. The synthetic voice could not pass the training procedure after several trials; we then manually removed the noise from the recordings using Audacity speech tool. We highlighted a small segment of the noise on the recording(s). In the Edit tab, we select the Noise Removal and Get Noise Profile as indicated in Figure 3.8:

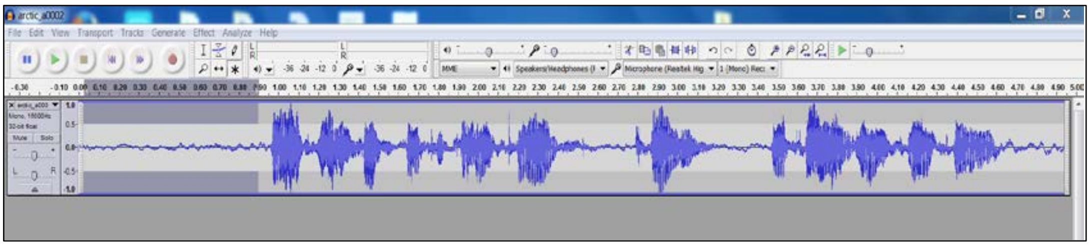


**Figure 3.8: The snapshot showing how to manually clean (removing noise and hiss on) an audio file using Audacity speech tool.**

After getting the Noise Profile, select the whole area of the audio file and set the parameters accordingly to remove the noise better, but do not over remove as you might reverberate the voice. All the audio files from the Tshivenda female were cleaned, and an example of a cleaned file is given in Figure 3.9. The cleaned training data was placed in a directory called *tshivenda_female*. To build this synthetic voice all 1131 files were used because one file was corrupted and discovered at a later stage and the speaker was unavailable for re-recordings.

After cleaning the recordings, the synthetic voice *tshivenda_female* was trained to completion. The HMMVoiceConfigure.speaker is set to *tshivenda_female*,

In the *voice-tshivenda_female-hsmm-5.2-SNAPSHOT-component.xml* file we then set:

name = “*tshivenda_female*”

description: “A Tshivenda speaking female English general voice."

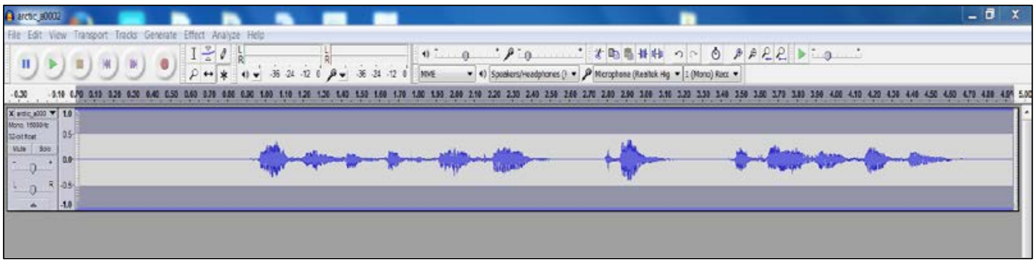


**Figure 3.9: The snapshot showing how the audio file looks after the noise and hiss is removed.**

### 3.6.3. *Voice sepedi_female*

The training data is placed in a directory called *sepedi_female*. The server was started first and during the execution of import we then set

HMMVoiceConfigure.speaker to *sepedi_female*. After training the synthetic voice, in the *voice/source/marytts/target/marytts-5.2-SNAPSHOT/download/* we edit the *voice-sepedi_female-hsmm-5.2-SNAPSHOT-component.xml*:

name = “*sepedi_female*”

description: “A Northern Sotho speaking female English general voice."

### 3.6.4. *Voice xitsonga_male*

During the development of the synthetic voice *xitsonga_male*, the only changes applicable to male synthetic voice development are effected in the settings:

In HMMVoiceConfigure Settings Editor set the following:

HMMVoiceConfigure.LowerF0 = 40

HMMVoiceConfigure.mgcBandWidth = 24 (for cepstral form)

HMMVoiceConfigure.mgcOrder = 24 (for cepstral form)

HMMVoiceConfigure.UpperF0 = 280

The above settings apply to all the male voice recommended by MARY TTS and the following settings apply only to this synthetic voice.

The HMMVoiceConfigure.speaker is set to *xitsonga_male* in this case and after the training phase the *voice-xitsonga_male-hsmm-5.2-SNAPSHOT-component.*xml is edited with:

name = “*xitsonga_male*”

description: “A Xitsonga speaking male English general voice."

### *3.6.5. Voice tshivenda_male*

The training speech data was placed in a directory called *tshivenda_male*. After a numerous prototyping of the voice, it is the better one. Once the training is complete we then set the name and description in the component.xml file as:

name = “*tshivenda_male*”

description: “A Tshivenda speaking male English general voice."

The installer.sh is used to add the new synthetic voice in the MARY TTS synthesis system.

### *3.6.6. Voice sepedi_male*

The training speech data was placed in a directory called *sepedi_male*. The server.sh and import.sh were executed and

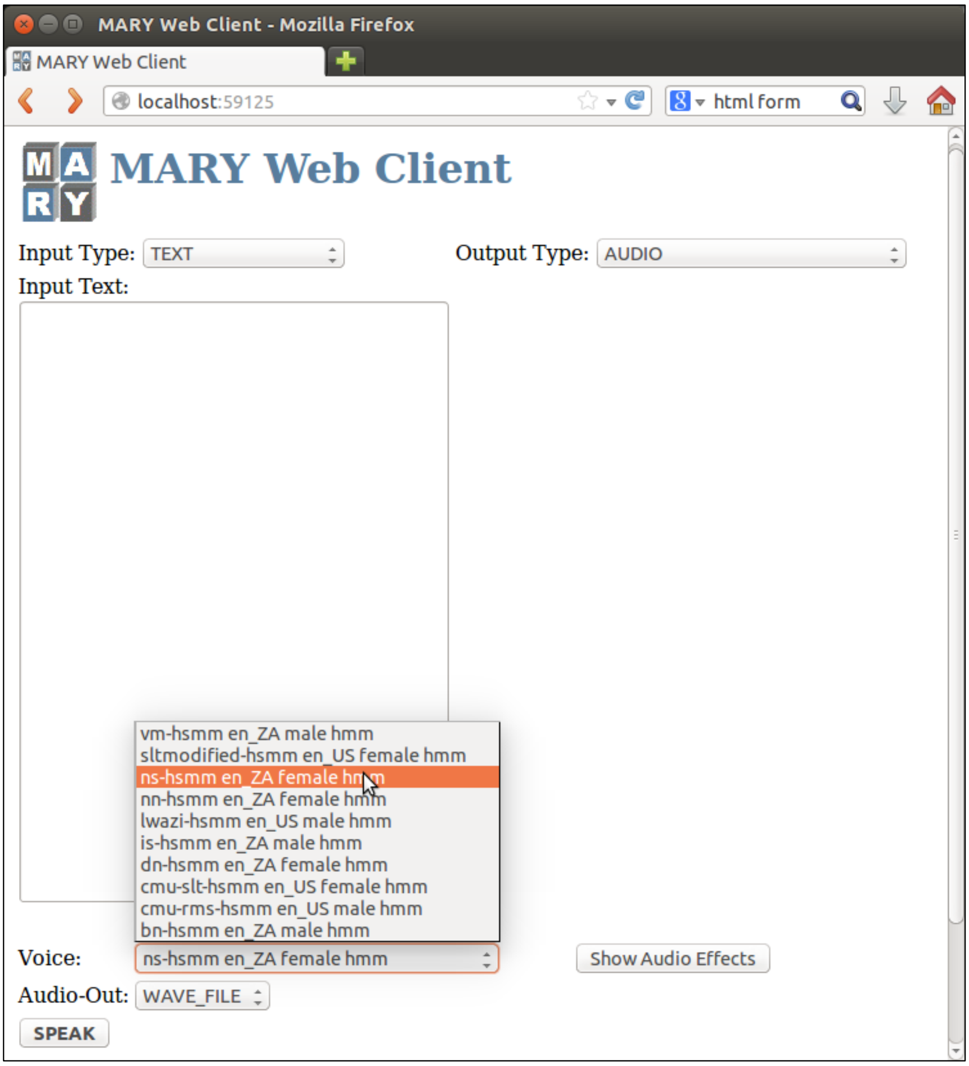
HMMVoiceConfigure.speaker was set to *sepedi_male*. After the training stage, we executed the copying.sh script and edited the voice-*sepedi_male*-hsmm-5.2-SNAPSHOT-component.xml file, setting the parameters to:

name = “*sepedi_male*”

description: “A Northern Sotho speaking male English general voice."

The installation was done and the server was restarted to effect the changes, as a result, all the synthetic voices were included in the system.

Figure 3.10 below shows all the ten synthetic voices developed in this study as they appear on the MARY Web Client. These voices include the four demonstration synthetic voices developed under the US English locale and the six synthetic voices developed under the ZA English locale.



Input Type: TEXT

Output Type: AUDIO

Input Text:

vm-hsmm en_ZA male hmm
sltmodified-hsmm en_US female hmm
ns-hsmm en_ZA female hmm
nn-hsmm en_ZA female hmm
lwazi-hsmm en_US male hmm
is-hsmm en_ZA male hmm
dn-hsmm en_ZA female hmm
cmu-slt-hsmm en_US female hmm
cmu-rms-hsmm en_US male hmm
bn-hsmm en_ZA male hmm

Voice: ns-hsmm en_ZA female hmm

Audio-Out: WAVE_FILE

**Figure 3.10: The MARY Web client interface snapshot consisting of the ten developed synthetic voices, four demonstration voices and six accented voices.**

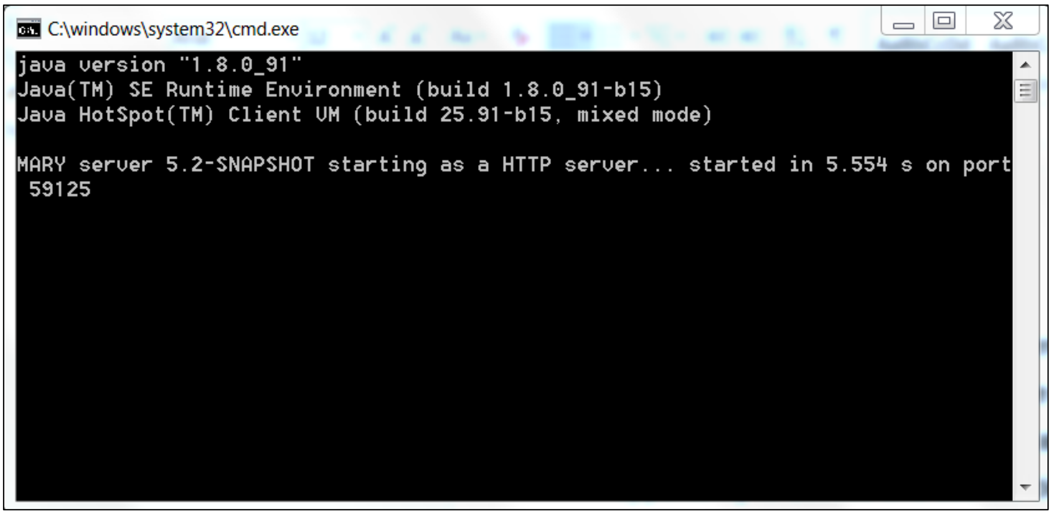
The University of Limpopo domain users can access all the developed synthetic voices on the MARY Web Client interface from their computers operating on Windows operating system. The access to the MARY Web Client was tested only on Mozilla Firefox browser, anyone attempting to access the voices is advised to do so using Mozilla Firefox. The computer used to develop the synthetic voices, and the MARY TTS server must be up and running. To access the MARY Web Client interface from Windows operating system the following steps must be execute:

- On Mozilla Firefox advanced options, select Network and open the Settings.
- Select the No proxy option from the given “Configure proxies to access the internet” options.
- After saving changes, enter the workstation’s internet protocol (IP) address (10.4.7.22) and the port (59125), i.e., 10.4.7.22:59125.

### 3.7. MARY Graphic User Interface (GUI) Client

The use of the MARY Web Client interface requires the user’s computer to be on the University of Limpopo domain and the MARY TTS server of the computer used to build the synthetic voices to be up and running at all times, which is not practical. The MARY GUI is meant to give Microsoft Window users the opportunity to access and use the developed synthetic voices at any time. MARY GUI client operates offline, unlike the MARY Web client interface. The following instructions are executed to access the MARY GUI client:

- Copy the marytts-5.2-SNAPSHOT.zip to your computer.
- Extract all to the same location (location is optional).
- Start the marytts-server by opening the *marytts-server.bat* file available on *marytts-5.2-SNAPSHOT/bin/.*
- Once the server is running, you can now open the MARY GUI client by opening the marytts-client.bat file also available on *marytts-5.2-SNAPSHOT/bin/*.
- All the ten voices developed in this research project will be available in the voice option.
- Change the output type to audio, to hear the synthesised speech.
- To save the audio file for later use, click the Save button and choose the wave (.wav) file type.

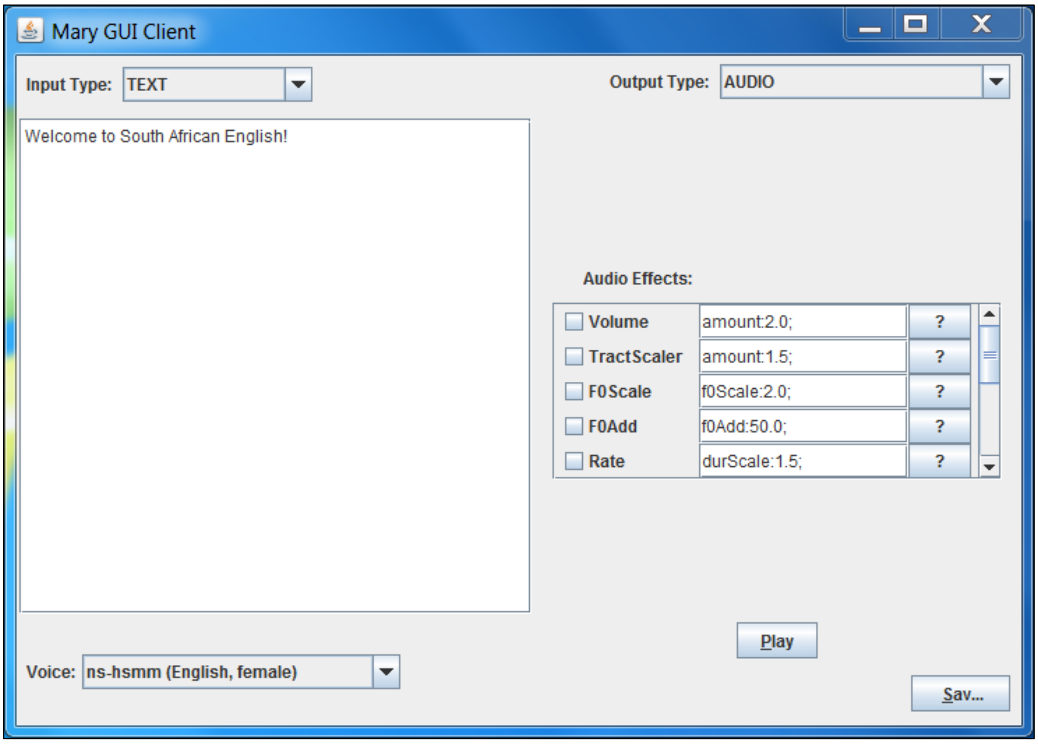


```
java version "1.8.0_91"
Java(TM) SE Runtime Environment (build 1.8.0_91-b15)
Java HotSpot(TM) Client VM (build 25.91-b15, mixed mode)

MARY server 5.2-SNAPSHOT starting as a HTTP server... started in 5.554 s on port
 59125
```

**Figure 3.11: The execution of the marytts-server Windows batch files to start the MARY server in Windows.**

Figure 3.11 shows a snapshot of a running MARY TTS server in Windows. After starting the server and opening the marytts-client the following interface in Figure 3.12 will appear. This GUI is the interface that will enable users to listen to the synthetic voices.



Mary GUI Client

Input Type: TEXT

Output Type: AUDIO

Welcome to South African English!

Audio Effects:

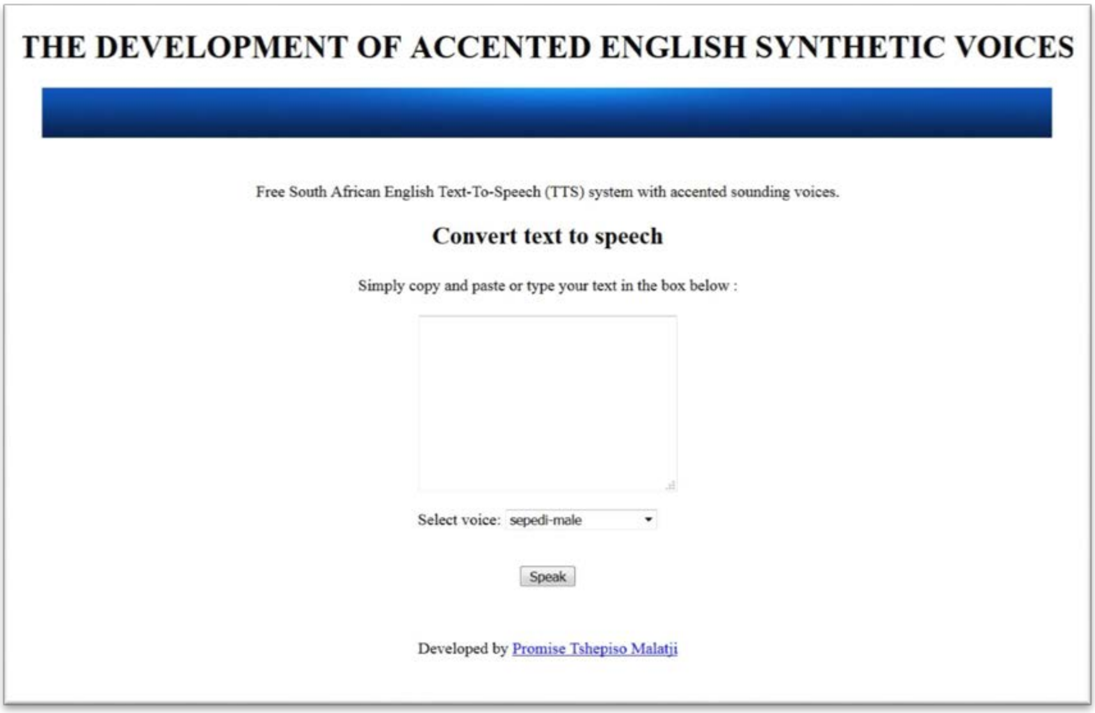
| | | |
|---|---|---|
| □ Volume | amount:2.0; | ? |
| □ TractScaler | amount:1.5; | ? |
| □ F0Scale | f0Scale:2.0; | ? |
| □ F0Add | f0Add:50.0; | ? |
| □ Rate | durScale:1.5; | ? |

Voice: ns-hsmm (English, female)

**Figure 3.12: The MARY GUI client interface for synthesising speech through the developed voices.**

## 3.8. Online Access to Synthetic Voices

The developed synthetic voices can also be accessed through the internet through: http://saenglishtts.byethost7.com. The website provides access to the six South African accented English synthetic voices developed in the study. The users are given the opportunity to listen to the synthetic voices anywhere at their own time and they can also download the audio files. Figure 3.13 shows the layout of the webpage to access the accented synthetic voices.



THE DEVELOPMENT OF ACCENTED ENGLISH SYNTHETIC VOICES

Free South African English Text-To-Speech (TTS) system with accented sounding voices.

Convert text to speech

Simply copy and paste or type your text in the box below :

Select voice: sepedi-male

Developed by Promise Tshepiso Malatji

Figure 3.13: The webpage to access the developed synthetic voices.

## 3.9. Summary

The MARY TTS engine was successfully installed and is working perfectly. The four demonstration synthetic voices were developed using readily available training speech data to test the functionality of the MARY TTS engine. In this study a total of six speakers were recruited and recorded reading the same text data. The training speech data used in the study is primary. After adding a South African English locale, six non-native accented English synthetic voices were developed using the created training speech corpus. A MARY GUI Client is provided for users to access the developed synthetic voices in Microsoft Windows. The developed synthetic voices can be accessed from any computer through the MARY GUI Client. A webpage is created to avail the developed accented synthetic voices to all users.

# CHAPTER 4

## 4. EVALUATION AND ANALYSIS

### 4.1. Introduction

In this chapter the researcher first looks at the method used to evaluate the developed synthetic voices. The three synthetic voice characteristics to be evaluated are discussed. The researcher also looks at the evaluation setup, the method used to select evaluators, the test data, and the criteria used to evaluate the developed synthetic voices. We provide the results and analysis of the synthetic voices evaluation in three classifications: per evaluator categories, per gender and the overall results. Lastly, the synthetic voice analysis using Praat speech analysis tool is given.

### 4.2. Evaluation Procedure

There are several methods of evaluating speech synthesis systems. The subjective listening tests are the most effective and popular method of evaluating TTS synthesis systems. In this study, evaluation was done through developing synthetic voices by exposing them to the potential end-users. We are evaluating three synthetic speech quality factors, namely, naturalness, intelligibility and similarity.

- Naturalness – relates to how close to human speech is the synthetic voice.
- Intelligibility – relates to how understandable is the speech uttered by the synthetic voice.
- Similarity – relates to how close is the synthesised speech to the original recordings.

In the naturalness and similarity tests, evaluators scored the speech produced by each synthetic voice on a 5–point scale. The averages of the scores obtained were calculated and compared. The mean of the scores is referred to as the mean opinion score (MOS). In intelligibility, the word error rate (WER) is calculated to determine the percentage of correct words the evaluators heard.

## 4.3. Evaluation Setup

This section looks at the method used to recruit evaluators, to select the test data and the data used per test conducted. The aspects considered when recruiting evaluators are stated below:

- Non-native speakers of English.
- Both male and female.
- High school learners, undergraduate students and postgraduate students and employees.
- On a volunteering basis.

The test data selection method was as follows:

- Sentences not in the training dataset, except for similarity test.
- Sentences were selected randomly from the Lwazi project TTS data.
- The number of sentences was fixed to eight sentences for naturalness test, because the ability of the synthetic voices to read different sentences separated by punctuation marks is testing. In the intelligibility test we are interested in the ability of evaluators to reproduce what is uttered by the synthetic voices hence we subject each synthetic voice to two sentences. Lastly, a similarity test compares each speaker's voice and the respective synthetic voice by using only one sentence per synthetic voice.

In this study, every evaluator was given a questionnaire (see Appendix F) that required their abridged biographic information and a participation consent form. Before the test was conducted, evaluators were informed that these tests are not recordings but a synthetic voice that is aimed at mimicking the human voice, and their judgement should be aligned with the nature of the work done. The evaluation was conducted in several sessions and on different days.

### *4.3.1. Naturalness Testing*

The evaluators were given the following test data which consisted of eight sentences to assist in fairly judging the TTS synthesis system. The sentences were then combined to form a paragraph to save evaluation time.

- data_001 “The machine knocks the chair aside, and keeps coming.”
- data_007 “You think it’s so easy.”
- data_008 “I can leave you out here, just like you left her.”
- data_009 “You got absolute zero.”
- data_018 “You owe me an explanation.”
- data_019 “He goes to the bedside table, and unscrews the earpiece.”
- data_020 “You were in, way over your head.”
- data_022 “Don’t let Longdale’s questionable choice of weapon give you any ideas.”

### *4.3.2. Intelligibility Testing*

In this test, evaluators were required to listen to two sentences uttered by each synthetic voice and after write what they heard from the synthesised speech. Unlike in the previous test, with intelligibility test the word error rate (WER) is used to determine the accuracy of the words heard by evaluators. We determine the WER by the following equation:

$$WER = \frac{100\ (S+D+I)}{N} \tag{3.1}$$

where *S* is the number of substitutions, *D* is the number of deletions, *I* is the number of insertions and *N* is the total number of words in the reference sentence.

The researcher used two sentences per synthetic voice to have a reasonable amount of words to be reproduced by evaluators. The two sentences used per synthetic voice are given below:

- Voice *xitsonga_female*

data_038 "They eat lunch at the snack counter."

data_039 "But after Mary was gone, that's when i got religious."

- Voice *tshivenda_female*

data_064 "It was the first real sharp collision of wills.”

data_065 "She was not by any means, the typical dowager."

- Voice *sepedi_female*

data_121 "As soon as he is in power, a change takes place."

data_122 "Why sir, how should it be otherwise."

- Voice *xitsonga_male*

data_056 "I always think he has the contrary to the evil eye."

data_057 "I hope she is fairly happy."

- Voice *tshivenda_male*

data_088 "He was a faithful and able minister of clement."

data_089 "My father sent him, his four prophetic verses."

- Voice *sepedi_male*

data_161 "They are always looking out for a reaction."

data_162 "These are things of which, we may well be proud."

### *4.3.3. Similarity Testing*

Unlike naturalness and intelligibility tests, this test required original recordings and their corresponding transcriptions and the synthesised speech for the evaluators to judge the similarity between the speakers’ and the synthetic voices. A different recorded wave file was used for each synthetic voice so that evaluators compare only the synthetic voice and the speaker’s recording not previously heard speech. Both the original recording and the synthesised speech were played, and evaluators were requested to score the similarity based on the 5–point Likert scale.

- Voice *xitsonga_female*

arctic_b0350 “Stand off butcher and baker and all the rest."

- Voice *tshivenda_female*

arctic_b0376 "Thought I, and a worthy fool he proved."

- Voice *sepedi_female*

arctic_b0391 "At sea, Tuesday, March 17, 1908."

- Voice *xitsonga_male*

arctic_b0368 "Please do not think that I already know it all."

- Voice *tshivenda_male*

arctic_b0381 "And how would we ever find ourselves."

- Voice *sepedi_male*

arctic_b0394 "The boy hesitated, then mastered his temper."

## 4.4. **Results and Analysis**

The developed synthetic voices were evaluated by 32 evaluators, with both genders evenly represented. The evaluators represent diverse groups (Sepedi, IsiZulu, SiSwati, and Xitsonga) of non-native speakers of English. The evaluation was not restricted to only the three indigenous languages used to collect data because the new synthetic voices are targeted at the indigenous language speakers. Three occupation levels of evaluators were represented in the study: school learners (in senior phase), undergraduate students and postgraduate students or employees. Eight evaluators were learners, with five males and three females. Sixteen undergraduate students were used, with the male and female representation of six and ten, respectively. Postgraduates/employees formed a collective of eight evaluators with five males and three females. The following results are based on the three categories: learners, undergraduates and postgraduates/employees for each synthetic voice.

The results in Section 4.4.1 to Section 4.4.9 compare the five synthetic voices (*xitsonga_female*, *xitsonga_male*, *tshivenda_male*, *sepedi_female*, *sepedi_male*) which were developed using the original recordings. Section 4.4.10 discusses the results of the only synthetic voice (*tshivenda_female*) which was developed with refined recordings.

### *4.4.1. Naturalness MOS per Occupation*

From Figure 4.1, we notice that learners had the lowest MOS for all the synthetic voices. Postgraduates/employees had the highest MOS for all the synthetic voices except for *xitsonga_female*. According to the postgraduates/employees, the synthetic voice *sepedi_female* is more natural than all the other voices with a MOS of 4.25. Learners are of the view that the synthetic *xitsonga_male* voice is unnatural with a MOS of 1.63. Both learners and undergraduates had the synthetic voice *sepedi_female* as their highest rated synthetic voice in naturalness.

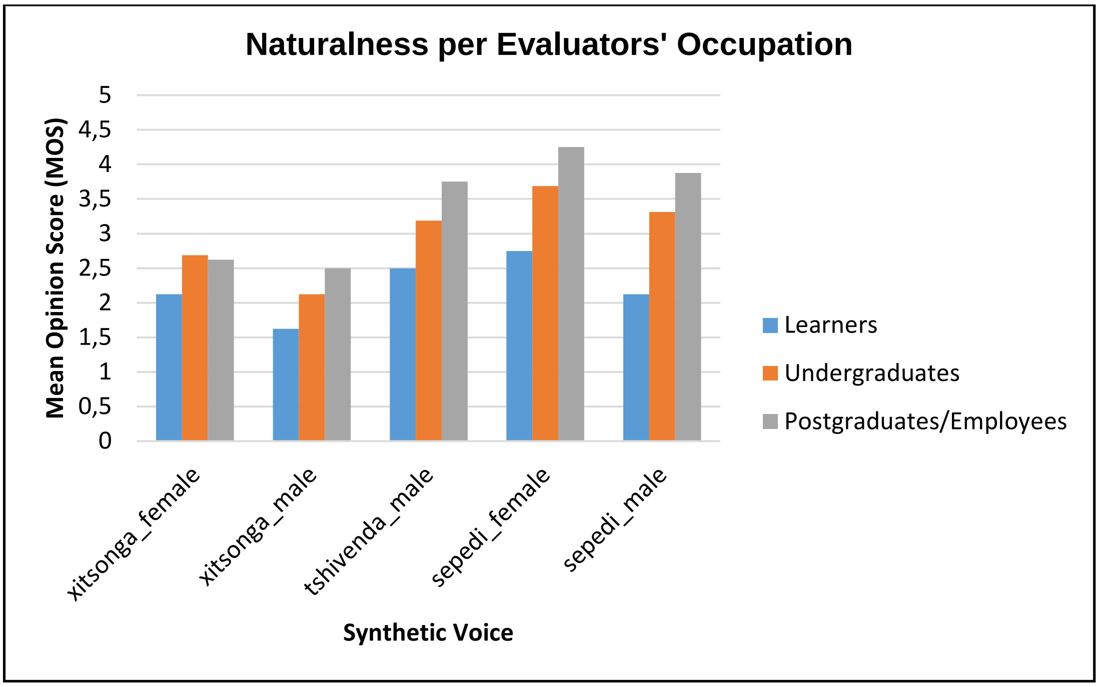


**Figure 4.1: The MOS of the naturalness of the five synthetic voices per occupation of evaluators.**

### *4.4.2. Intelligibility WER per Occupation*

Figure 4.2 shows the variation in intelligibility within the five synthetic voices. The learners had the highest WER for all the synthetic voices. The synthetic *sepedi_female* voice is unintelligible to learners with the highest WER of 88.89%. Postgraduates/employees had the lowest WER of 22.92% for the synthetic *sepedi_male* voice. A total of 77.08% on the words uttered by synthetic *sepedi_male* voice were understood by postgraduates/employees. The most intelligible synthetic voice according to undergraduates is also *sepedi_male*, with 75% of its words correctly captured. Learners and undergraduates hardly heard a word spoken by the synthetic voices *xitsonga_male* and *sepedi_female* with a WER of 72.79% for both, whereas postgraduates/employees had a tough time listening to the synthetic *sepedi_female* voice with a WER of 63.89%. From the results we note that occupation has an impact on intelligibility.

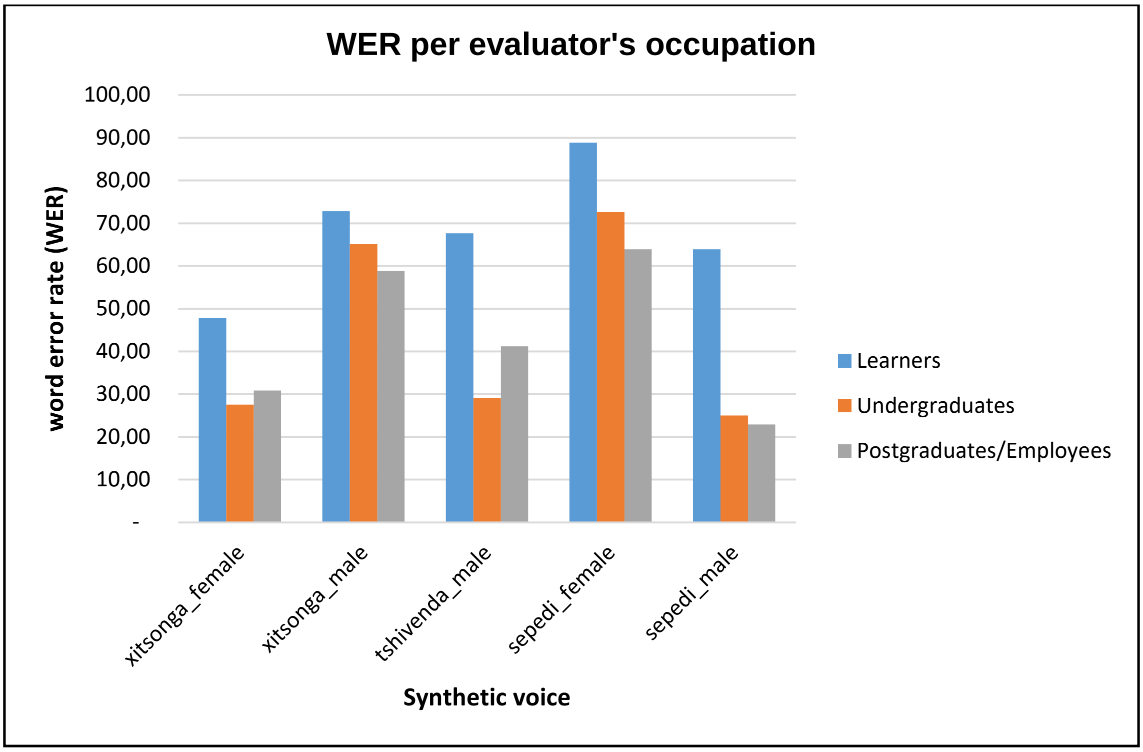


**Figure 4.2: The WER of intelligibility of the five synthetic voices per occupation of evaluators.**

### *4.4.3. Similarity MOS per Occupation*

Most of the synthetic voices created were said to be almost similar to their respective speakers by postgraduate/employees, except for the synthetic *xitsonga_male* voice. According to postgraduates/employees the synthetic *tshivenda_male* voice sounds like the speaker with the MOS of 4.25. From figure 4.3 a collective total of 75% of the postgraduates/employees said that the synthetic *tshivenda_male* voice sounds like the speaker or sounds totally like the speaker. The MOS by learners was the lowest for all the synthetic voices except for the synthetic *sepedi_female* voice with 3.50. Undergraduates had the only highest MOS of 3.44 amongst the other two categories on the synthetic *sepedi_male* voice although there is no significant difference between them and the postgraduates/employees.

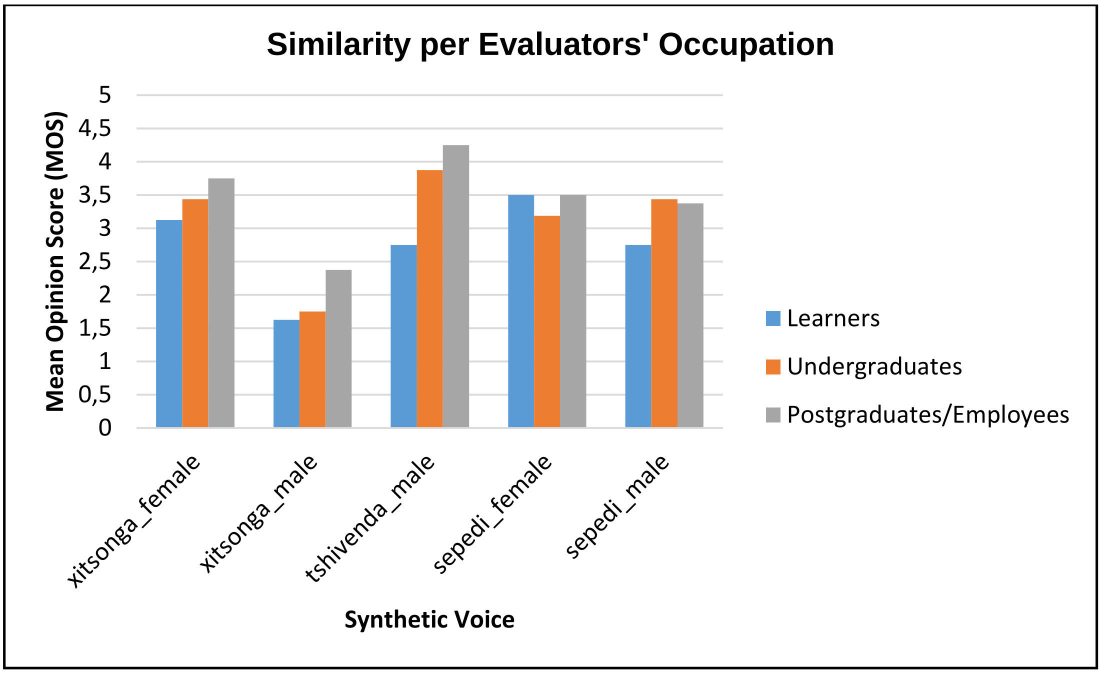


**Figure 4.3: The MOS of similarity of the five synthetic voices per occupation of evaluators.**

Based on the results per evaluators' occupation, it is noted that learners scored the synthetic voices low on all three tests and as such this negatively affected the overall MOS of the synthetic voices. Undergraduate students had an average MOS of the three categories whereas the postgraduates/employees had the highest MOS in most cases. There were few concerns noted from the evaluation by learners, such as, their inability to construct English sentences.

### *4.4.4. Naturalness MOS per Gender*

Figure 4.4 shows the results of the naturalness test based on the evaluators' gender. According to the female evaluators, the synthetic *sepedi_female* voice is natural with a MOS of 3.73. Although the MOS for the same synthetic *sepedi_female* voice is 3.09 for male evaluators, it remains the highest MOS for males. It is noted that the synthetic *xitsonga_male* voice had the lowest MOS from both genders. Female evaluators had the highest MOS in all the synthetic voices even though both genders were evenly represented. A combined total of

81.25% of male evaluators said the synthetic *xitsonga_male* voice was unnatural or completely unnatural.

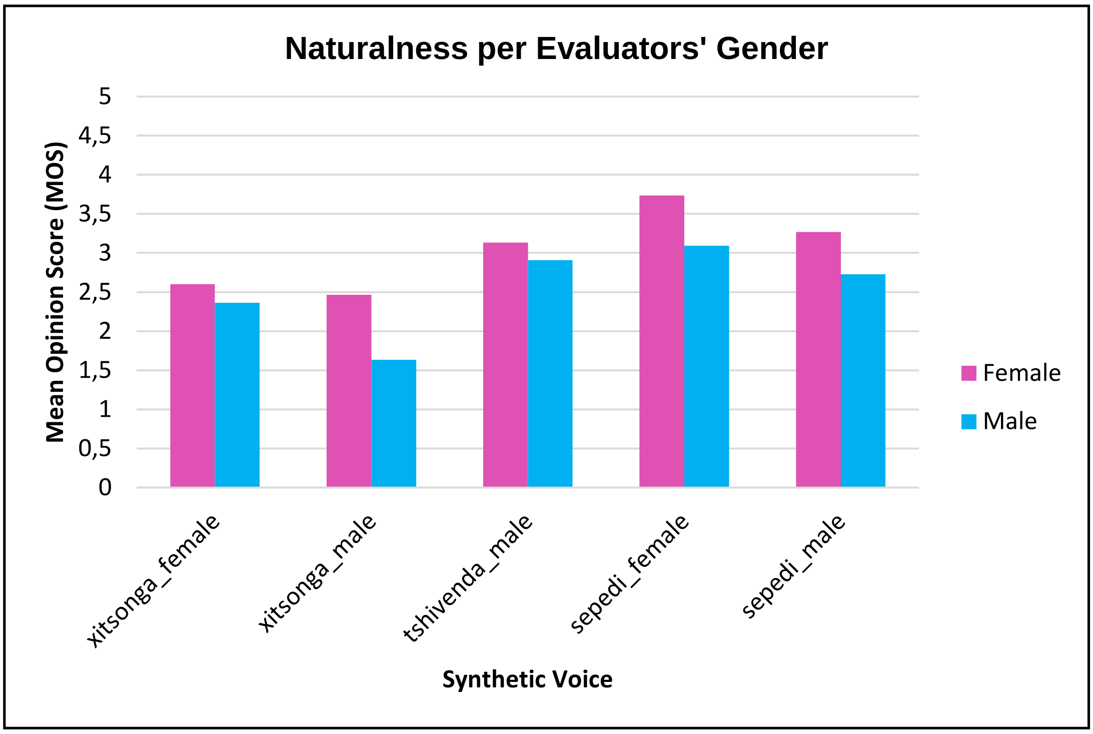


**Figure 4.4: The MOS of the naturalness of the five synthetic voices per gender of evaluators.**

### *4.4.5. Intelligibility WER per Gender*

From Figure 4.5 it is noted that female evaluators were able to reproduce the uttered words than male evaluators. In all the synthetic voices, females have a lower WER, with the lower WER of 27.08%. An approximation of 73% of the words spoken by both synthetic voices *xitsonga_female* and *sepedi_male* were intelligible to female evaluators. The lowest WER for male evaluators is 39.71%; meaning males were able to correctly write 60.29% of what was articulated by the synthetic *xitsonga_female* voice. The synthetic *sepedi_female* voice has the highest WER for both genders; it remains the unintelligible synthetic voice in terms of gender evaluation.

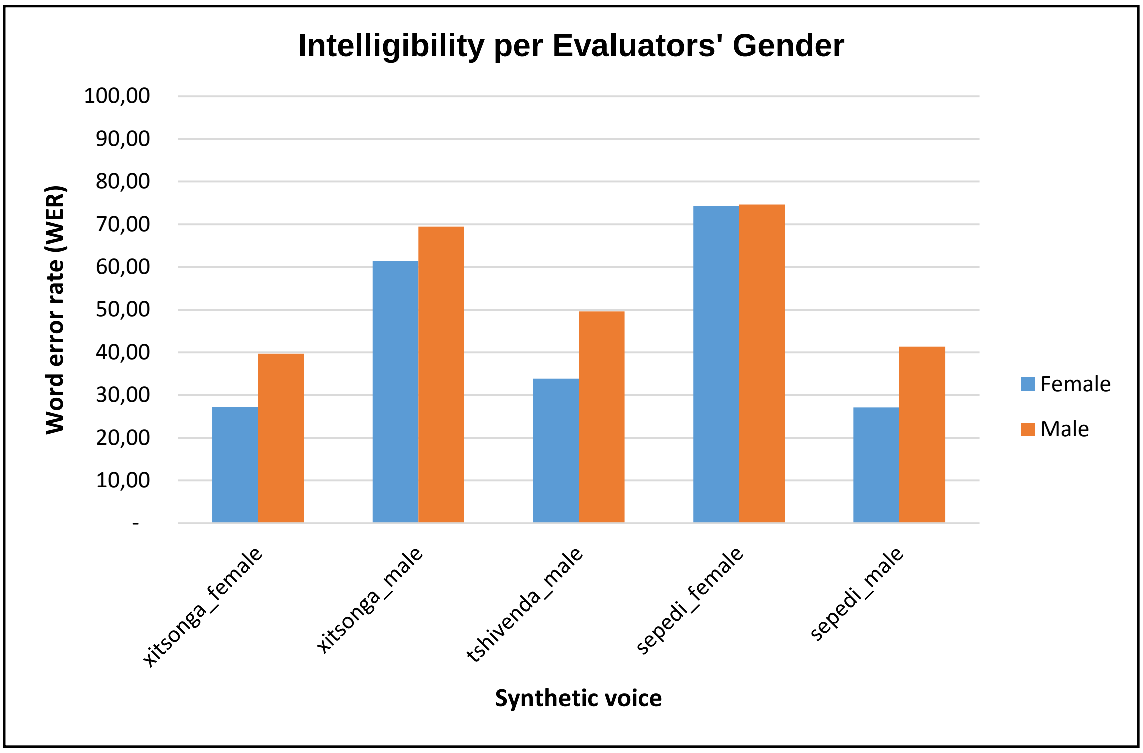


**Figure 4.5: The WER of intelligibility of the five synthetic voices per gender of the evaluators.**

### *4.4.6. Similarity MOS per Gender*

Figure 4.6 shows the results according to the gender of evaluators. With a MOS of 3.87, female evaluators are of the view that the synthetic *tshivenda_male* voice is similar to the speaker while male evaluators believe that both the synthetic voices *xitsonga_female* and *sepedi_female* are similar to their respective speakers with a MOS of 3.46 each. The females had a higher MOS for all the synthetic voices except for *sepedi_female*. According to female and male evaluators, the synthetic *xitsonga_male* voice sounds like a different speaker or sounds like a totally different speaker with a combined total of 62.50% and 87.50% evaluators, respectively.

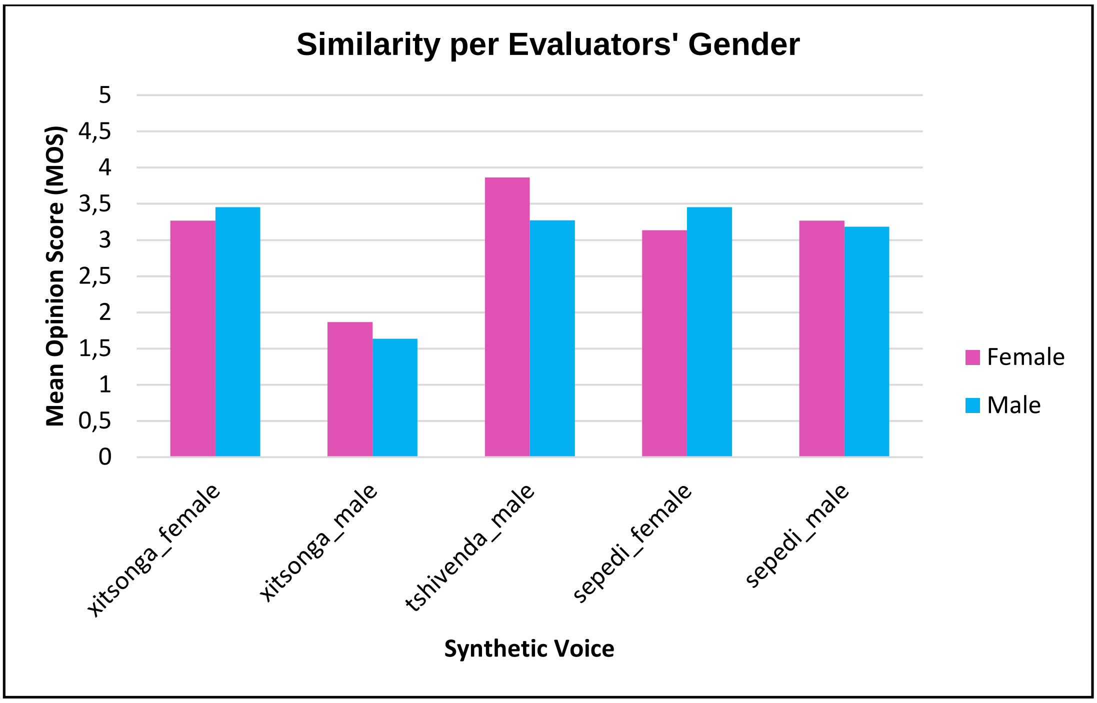


**Figure 4.6: The MOS of similarity of the five synthetic voices per gender of evaluators.**

From the evaluation based on gender, it is conclude that females had a higher MOS than males in all the tests except for the synthetic *sepedi_female* voice in similarity test. The possible cause of the imbalance in the MOS might be the fact that females form a bigger fraction of the undergraduate students while males form the greater portion of the learners who had the lowest MOS in all the tests.

To determine the general overview about the synthetic voices the researcher then computed the overall MOS of all the evaluators for the two tests, naturalness and similarity. The general WER for intelligibility test was also computed. Section 4.4.7 to 4.4.9 discusses the overall results obtained per test.

### *4.4.7. Overall Naturalness MOS*

The total MOS for all the evaluators for naturalness indicates that more than half of the synthetic voices are above average, with the synthetic *sepedi_female* voice being almost natural at 3.59 MOS. From figure 4.7 we note that the synthetic voice *xitsonga_male* was scored the least with MOS of 1.84 by all the evaluators.

Both synthetic voices *sepedi_female* and *sepedi_male* are said to be completely natural by 15.63% and 12.50% of the evaluators, respectively. The synthetic *sepedi_female* voice had the highest total percentage of 56.25% of the evaluators saying it is natural or completely natural. A total of 28.16% of evaluators said that the synthetic *xitsonga_male* voice is completely unnatural.

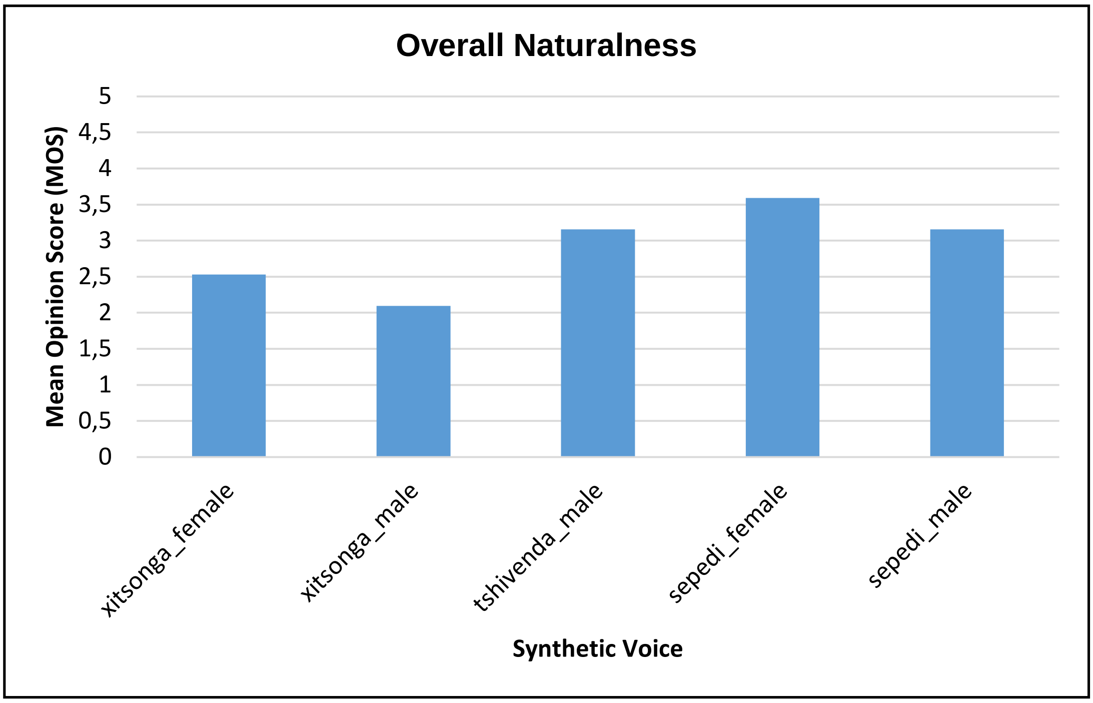


**Figure 4.7: The overall MOS of the naturalness of the five of the synthetic voices.**

### *4.4.8. Overall Intelligibility WER*

Generally, the synthetic *sepedi_female* voice is the worst intelligible synthetic voice developed with only 25.52% of its utterances understood by evaluators. Even though the synthetic *sepedi_female* voice has a higher MOS in the overall naturalness test, it remains the worst intelligible synthetic voice. From figure 4.7 we note that there is no significant difference between the intelligibility of the synthetic *xitsonga_female* voice and the synthetic *sepedi_male* voice with a WER of 33.46% and 34.20%, respectively. The evaluators were able to correctly write

76.54% of the words spoken by the synthetic voice *xitsonga_female*. The synthetic voice *tshivenda_male* dominated the similarity tests above but only 58.27% of the words it utters are intelligible.

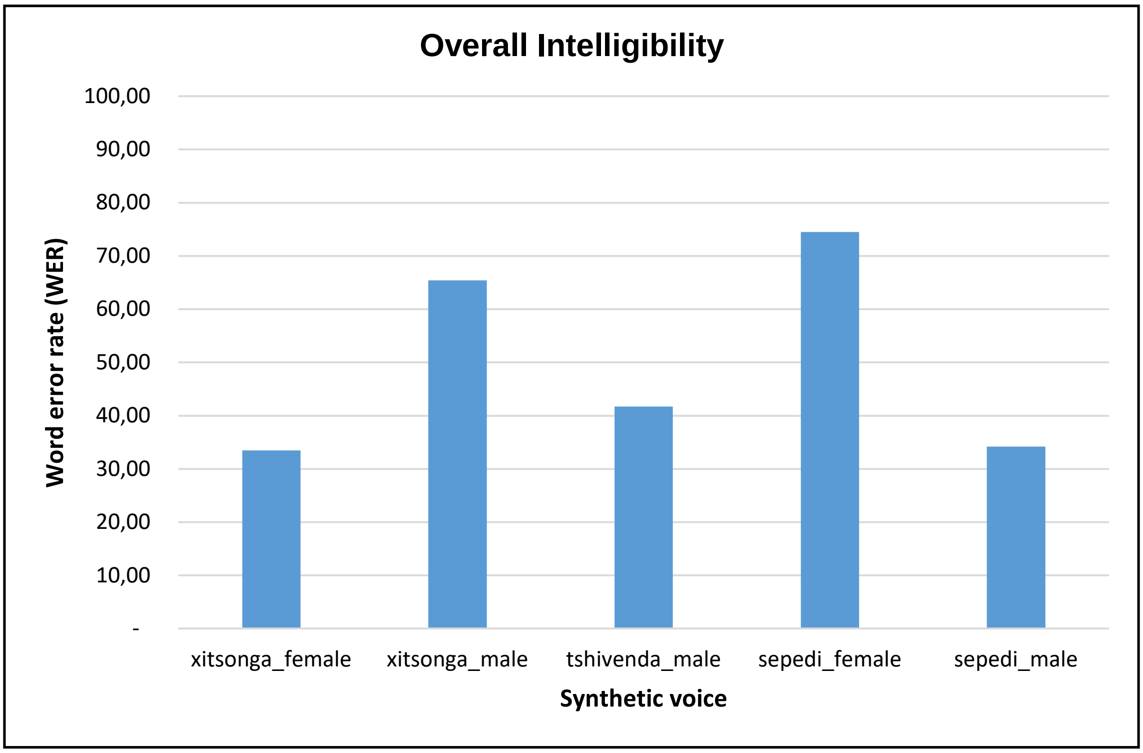


**Figure 4.8: The overall WER of the five synthetic voices.**

### *4.4.9. Overall Similarity MOS*

The synthetic *tshivenda_male* voice sounds similar to the speaker's voice than all the other voices. From the evaluators, 28.13% said that the synthetic *tshivenda_male* voice sounds exactly like the speaker. With a collective total of 62.50% of the evaluators saying that the synthetic *tshivenda_male* voice sounds like the speaker or sounds exactly like the speaker. Three other synthetic voices (*xitsonga_female*, *sepedi_female* and *sepedi_male*) sound almost like their respective speakers with the MOS of above 3. The only synthetic voice that sounds like a different speaker was the synthetic *xitsonga_male* voice with a

MOS of 1.88. Twenty-five evaluators scored the synthetic voice *xitsonga_male* a maximum score of 2.

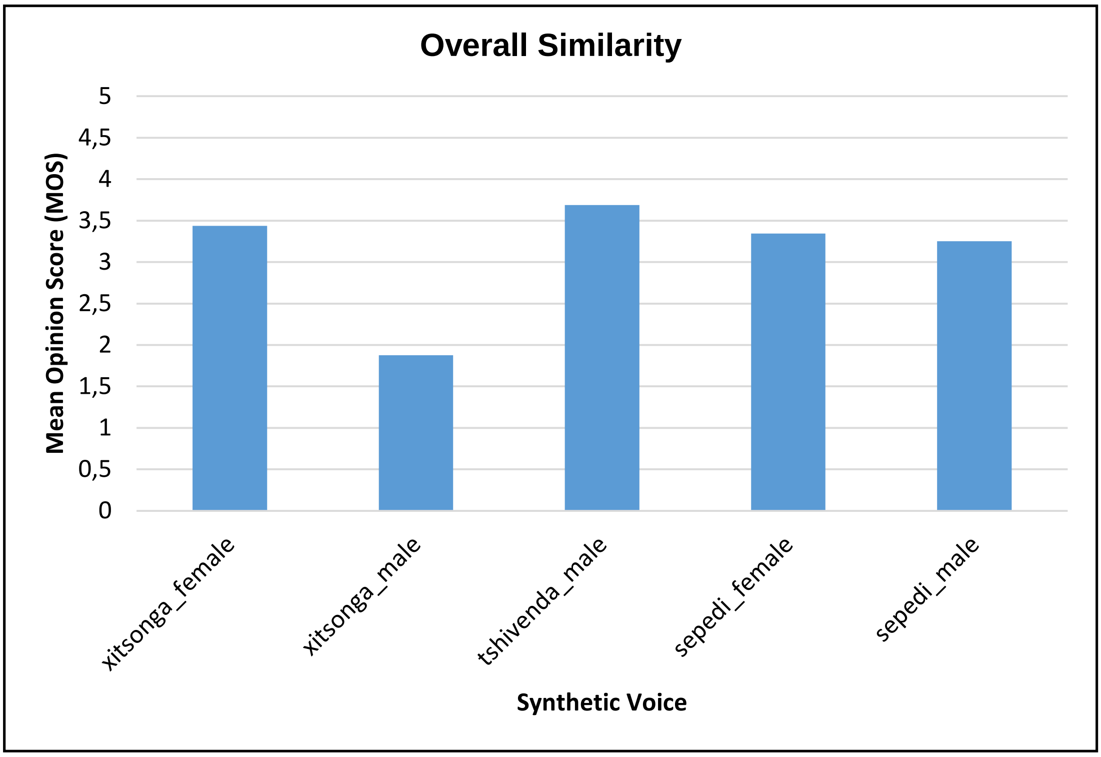


**Figure 4.9: The overall similarity MOS for the five synthetic voices.**

From the three overall Figures 4.7, 4.8 and 4.9, it is noted that the synthetic voices have different strengths. None of the synthetic voices has a greater score in all the tests. According to the evaluators, the synthetic *sepedi_female* voice sounds natural than the other synthetic voices. The synthetic *xitsonga_female* voice is considered the intelligible of the five synthetic voices, although there is no significant difference between its intelligibility and that of the synthetic *sepedi_male* voice. Regarding similarity, *tshivenda_male* is the synthetic voice that scored a higher MOS. All the synthetic voices had a MOS of less than four in all the tests even though some synthetic voices had a MOS of more than four in evaluation per occupation. There are several possible contributing factors to the quality of the synthetic voices developed, namely:

- The use of unprofessional speakers to collect training speech data.
- The use of an office for recording instead of a professional studio.
- Poor quality microphone for recording.

### *4.4.10. Refined Synthetic Voice Results*

The results of the synthetic *tshivenda_female* voice are not impressive. In Figure 4.10 the researcher looked at the results per evaluators' occupation. Both learners and undergraduate students had a naturalness MOS of less than 2 in Figure 4.10(A). No category of evaluators said that the synthetic *tshivenda_female* voice is natural. Figure 4.10(B) indicates positive results for the synthetic *tshivenda_female* voice. According to undergraduates and postgraduates in figure 4.10(B), the synthetic *tshivenda_female* voice is almost similar to the speaker with a MOS of 3.56 and 3.88, respectively. Figure 4.10(C) shows that all the learners were unable to hear a single word uttered by this synthetic voice. Postgraduates could only hear 45.83% of the words synthesised by synthetic *tshivenda_female* voice.

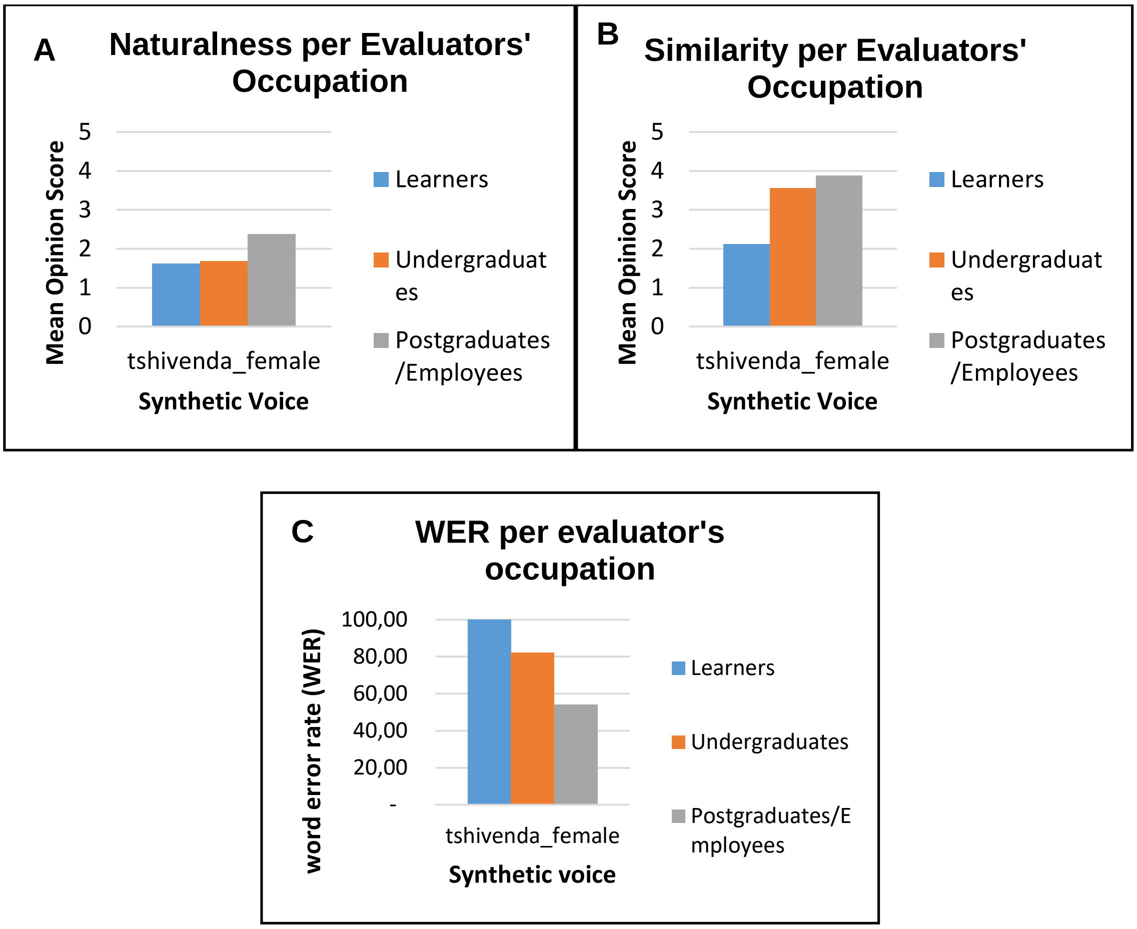


**Figure 4.10: The graphical representation of the evaluation results of the synthetic voice tshivenda_female based on occupation of evaluators.**

Figure 4.11 graphically represents the results of the evaluation based on gender for the synthetic *tshivenda_female* voice. In Figure 4.11(A) we note that females and males had a maximum MOS of 2 and 1.36, respectively. According to the two groups of evaluators, the synthetic *tshivenda_female* voice is unnatural. Females are slightly of the impression that the synthetic *tshivenda_female* voice similar to the speaker with a MOS of 3.4. There is no significant difference between the WER of females and males, with 77.78% and 81.60%, respectively.

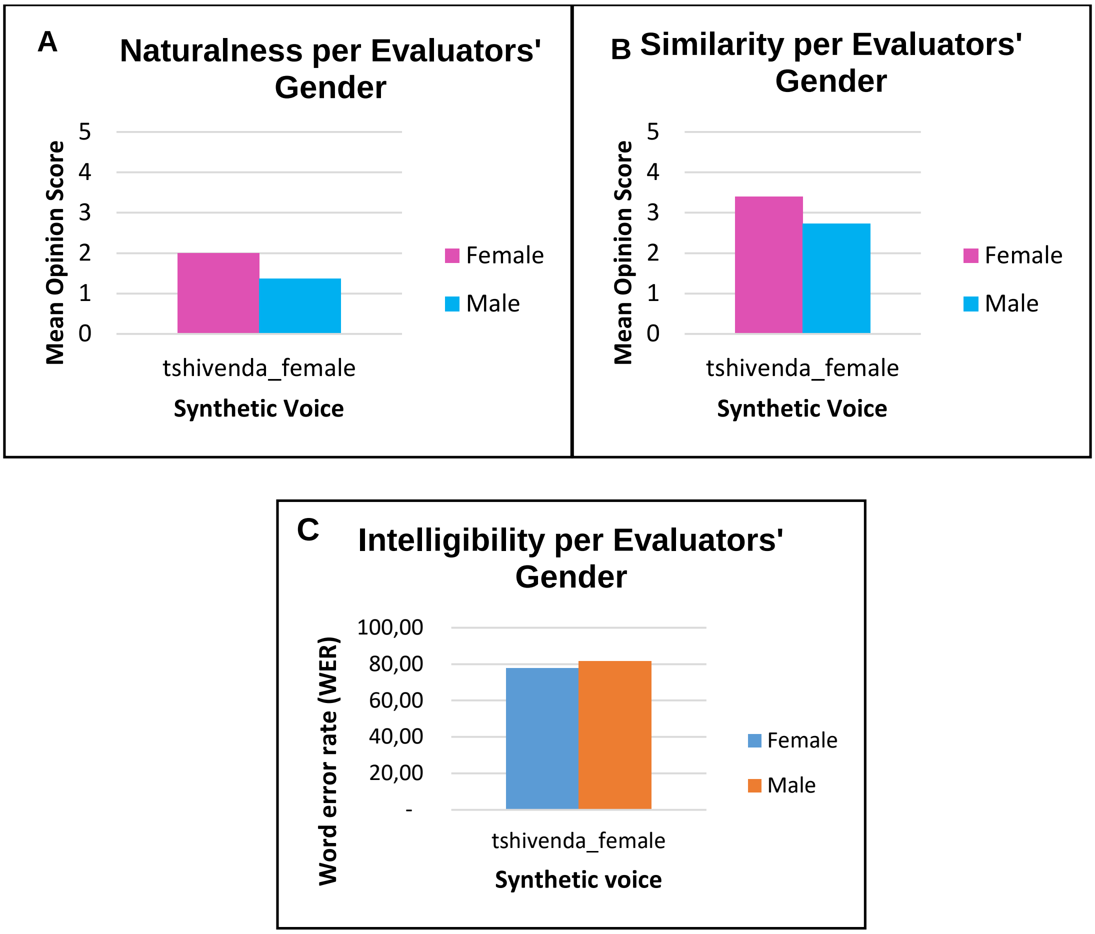


**Figure 4.11: The gender-based results for all the three tests conducted on the synthetic voice tshivenda_female.**

Looking at the overall results of the synthetic *tshivenda_female* voice in terms of naturalness, the researcher notes that generally the synthetic *tshivenda_female* voice is said to be unnatural with a MOS of 1.84. Figure 4.12.B shows that there is a level of similarity between the synthetic voice and the speaker. Generally, only 20.31% of the words synthesised are intelligible. Figure 4.12 gives a graphical structure of the results.

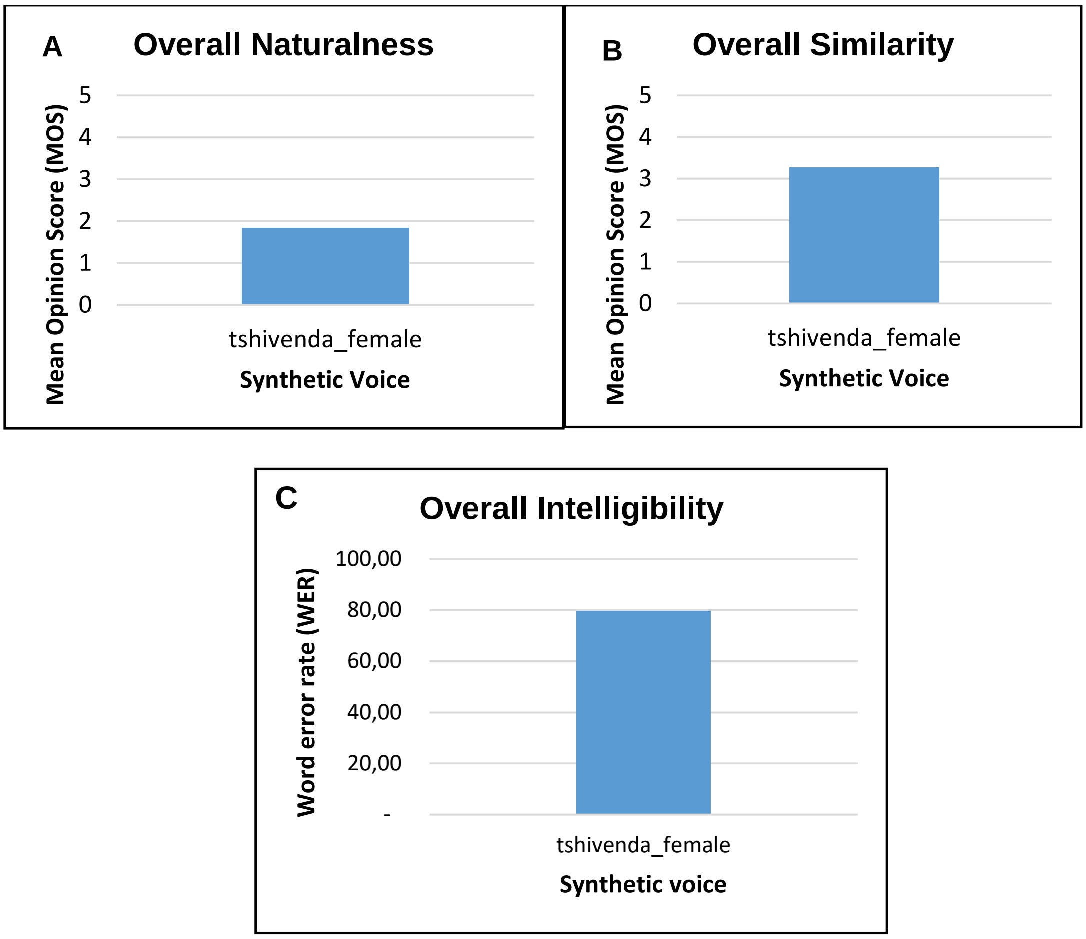


**Figure 4.12: The graphical representation of the overall results for the three tests for the synthetic voice tshivenda_female.**

## 4.5. Voice Analysis using Praat

In this section we used Praat speech tool to compare our synthetic voices with the respective recordings. Figure 4.13 shows the processing modules involved in speech synthesis and also how data flows from input as text to intermediate result which is speech. The speech produced in this TTS synthesis system serves as input to the speech analyser to compare with the recording from the researcher's speech database.

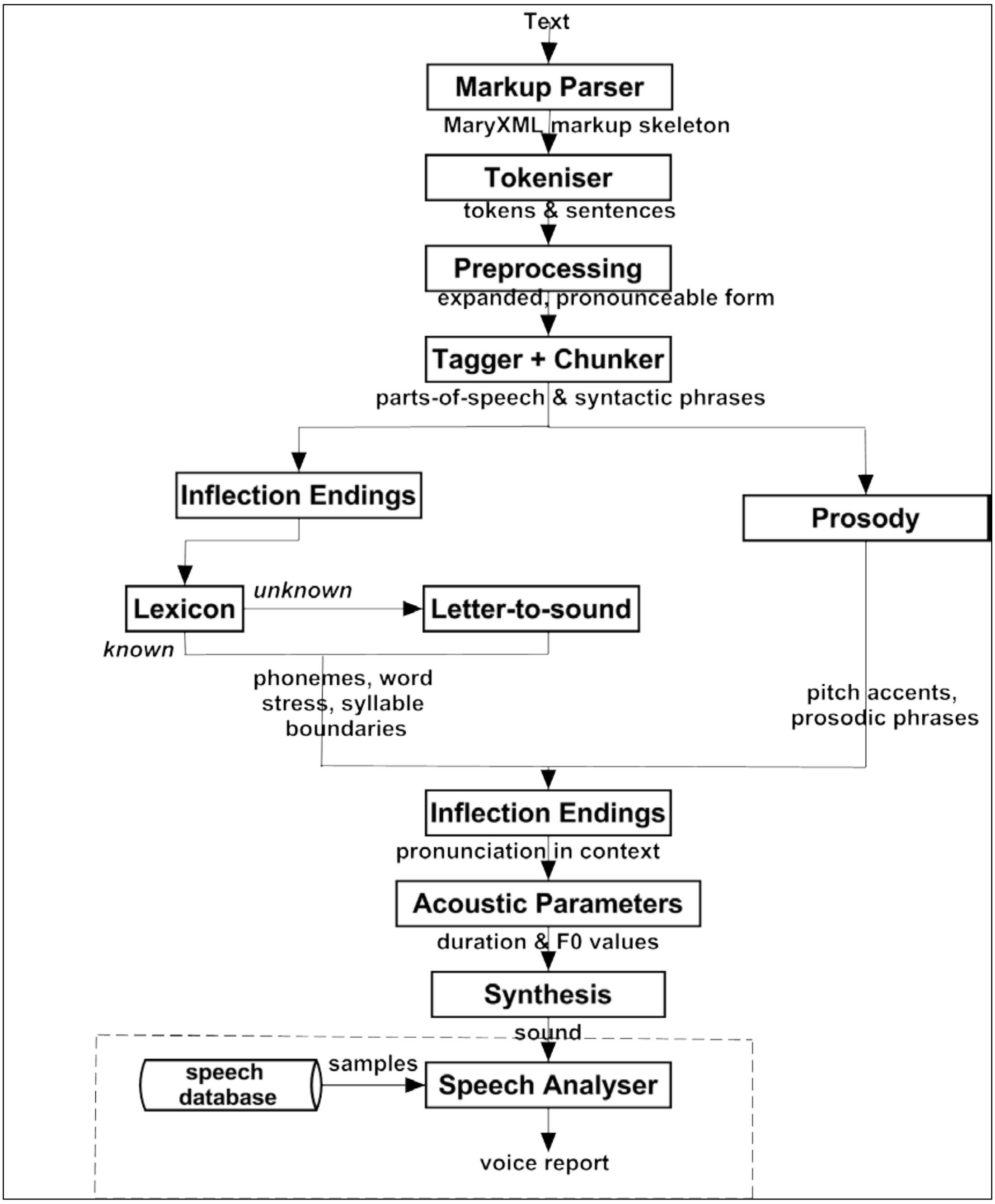


**Figure 4.13: The flow diagram of speech synthesising phase with the usage of Praat speech tool to compare speech sounds, (adapted from Schröder. & Trouvain 2001).**

Appendix E shows the voice reports for all the synthetic and speaker voices. Each synthetic voice used the same sentence used by the speaker. Based on Figure 4.14 the synthetic *xitsonga_female* voice has approximately the same pitch standard deviation as the original recording of the speaker, detailed in Appendix E. There is a significant difference between the synthetic *xitsonga_male* voice and the original recording of more than 60 hertz (Hz) in the maximum pitch. The synthetic *tshivenda_male* voice has no significant difference in terms of pitch with the speaker's voice pitch with a standard deviation of less than 4 Hz. Based on the standard deviations in Figure 4.14, the synthetic voice that has the same pitch as the original speaker is *xitsonga_female* with the difference of 2.60 Hz (36.01 Hz - 33.41 Hz). According to the evaluators the synthetic *tshivenda_male* voice is the most similar synthetic voice to its speaker whereas in terms of pitch, the synthetic *xitsonga_female* voice is the most similar one. The speakers' voices have less breaks than the synthetic voices except for synthetic *xitsonga_female* voice. The difference in the degree of voice breaking is more significant in the synthetic voices *tshivenda_female* and *sepedi_female* with 23.23% and 21.08%, respectively. The Figure shows the standard deviation between the pitch of each synthetic voice and its respective speaker's.

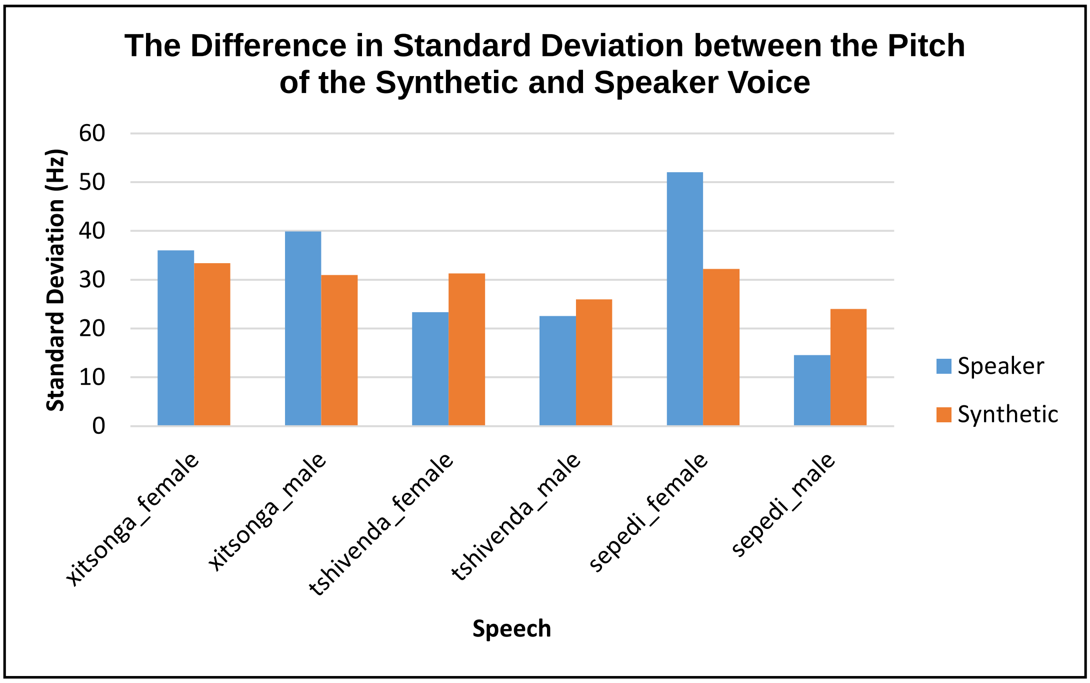


**Figure 4.14: Graphical representation of the deviation of the synthetic voice pitch from the speaker's pitch for each synthetic voice.**

## 4.6. **Summary**

This chapter gave an analysis of the results based on the three categories: per evaluator occupation, per evaluator gender and overall. It was noted that the MOS of all the synthetic voices in the overall analysis was low, and the possible causes were indicated. In the evaluator occupation, learners had the low MOS for almost all the synthetic voices in all the tests. The possible causes for males to have low MOS in all synthetic voices except for the synthetic *sepedi_female* voice in similarity test have been discussed. The possible contributing factors to the quality of the developed synthetic voices have also been outlined.

# CHAPTER 5

## 5. CONCLUSION AND FUTURE WORK

The research study presented a way of creating and developing new accented English synthetic voices using the open source speech processing toolkits. The TTS synthesis system developed was extended to include four demonstration voices from native speakers of English. The inclusion of the synthetic demonstration voices helped in determining the size of text corpus used in developing the training speech data. The TTS synthesis system developed showed that the synthetic voices have the potential of being more intelligible, natural and similar to the respective speakers.

### 5.1. **Conclusion**

The results obtained from subjective listening tests conducted show that some of the developed synthetic voices are fairly natural, understandable and similar to the respective speakers. The evaluation results as per evaluators' category indicate that there exists an increasing trend in the MOS from learners to postgraduates/employees. While postgraduates/employees had a high MOS, learners scored the synthetic voices the lowest in all tests. The undergraduate had an average MOS of the other two categories. In the evaluation per gender, females had a high MOS for all the synthetic voices except one. The low scoring of learners in all the tests for all the synthetic voices lowered the overall MOS. The factors that contributed to the low scoring, especially for intelligibility test by learners, were noted in chapter 4. The developed synthetic voices were more accepted by university students and employees than by high school learners. The overall MOS suggests that there is room for improvement in this research. The developed synthetic voices had many evaluators interested in the study and enjoyed listening to the synthesised speech.

## 5.2. **Future Work**

To improve the quality of the developed synthetic voices, the following amongst others must be considered:

- The use of professional speakers in the development of training speech data is very critical. Using unprofessional speakers leads to more data collection time and training speech data with different speaking rate, tone and pitch. Professional speakers will eliminate the random pauses on sentences and violation of punctuation marks.
- Speech recording should be done in a proper professional studio where the noise level is minimal, and the environment is conducive for such an activity. The use of an office as recording space yields unnecessary noise on the speech and also caused fatigue to speakers. This fatigue resulted in many speakers withdrawing from the study and consequently prolonging the data collection process because new speakers needed to be recruited. The setup in a work office is not comfortable for speakers.
- In line with a proper recording studio, good quality microphones are very crucial to obtain the best speech data. These microphones will eliminate the hiss or noise caused by the low-quality microphones used in this study.
- Where possible, it is vital to avoid noise removal from the training speech data. As this has the potential to cause the speaker’s voice reverberate, and that will negatively affect the quality of the synthetic voice.
- The use of a powerful computational computer that operates on GPU will help in improving the quality of the developed synthetic voices. Such a computer will enable the training of the synthetic voices through the DNNs to produce state-of-the-art synthetic voices.

The use of MARY TTS to develop synthetic voices has the potential to produce better quality synthetic voices provided the training speech data is clean and of good quality.

# REFERENCES

Baloyi, N., 2012, *A Text-to-Speech Synthesis System for Xitsonga using Hidden Markov Models*, Unpublished Masters of Science Mini-Dissertation, Department of Computer Science, University of Limpopo, viewed 29 February 2016, from http://ul.netd.ac.za/bitstream/handle/10386/1021/Baloyi_n_2012.pdf?sequence=1&isAllowed=y.

Barnard, E., Davel, M.H., van Heerden C., de Wet F. & Badenhorst, J., 2014, 'The NCHLT Speech Corpus of the South African Languages', *SLTU-2014*, St. Petersburg, Russia, May 2014. 194 – 200.

Black, A.W, *Festvox*, cmu_us_slt_arctic-0.95-release.tar.bz2, viewed 01 February 2016, from http://festvox.org/cmu_arctic/cmu_arctic/packed/cmu_us_slt_arctic-0.95-release.tar.bz2.

Black, A.W., *Festvox*, festival-2.4-release.tar.gz, viewed 30 January 2016, from http://festvox.org/packed/2.4/festival-2.4-release.tar.gz.

Black, A.W., *Festvox*, festvox-2.7.0-release.tar.gz, viewed 30 January 2016, from http://festvox.org/festival-2.7/festvox-2.7.0-release.tar.gz.

Black, A.W., *Festvox*, speech_tools-2.4-release.tar.gz, viewed 30 January 2016, from http://festvox.org/packed/festival/2.4/speech_tools-2.4-release.tar.gz.

Black, A.W., Zen, H., & Tokuda, K., 2007, 'Statistical Parametric Speech Synthesis'*, Proceedings of ICASSP*, 2007, 1229 – 1232.

Boersma, P., 2017, *Phonetic Sciences, Amsterdam*, praat6014_linux64.tar.gz, viewed 30 January 2017, from http://www.fon.hum.uva.nl/praat/praat6014_linux64.tar.gz.

Boersman, P. & Weenink, D., Praat: doing phonetics by computer, viewed 18 November 2015 from http://www.fon.hum.uva.nl/praat/.

Bułojčyk, A., 2015, [mary-users] *New Voice Creation*, viewed 01 March 2016, from http://www.dfki.de/pipermail/mary-users/2015-November/001779.html.

Cambridge University Engineering Department (CUED) Machine Intelligence Laboratory, *HTK3*, HTK-3.4.1.tar.gz, viewed 30 January 2016, from https://htk.eng.cam.ac.uk/ftp/software/HTK-3.4.1.tar.gz.

Clopper, C. G., & Bradlow, A. R., 2009, ‘Free classification of American English dialects by native and non-native listeners’, *Journal of Phonetics*, 436 – 451.

Code Welt Speak online text to speech synthesizer, (Code welt.com), (Last updated: 10/03/2014), viewed 17 August 2015 from http://codewelt.com/proj/speak.

Coto-Jiménez, M., Goddard-Close, J., 2016, ‘Speech Synthesis Based on Hidden Markov Models and Deep Learning’, *Research in Computing Science 112 (2016)*, 19 – 28.

Echo productions radio & television advertising: South African television channel, viewed 03 December 2015, from http://www.echoads.co.za/index.php/articles/59-south-africas-television-channels-.

Ghorshi, S., Vaseghi, S., & Yan, Q., 2008, ‘Cross-entropic comparison of formants of British, Australian and American English accents’, *In Speech Communication*, Elsevier 2008, 50, 564 – 579.

GitHub, 2017, New Voice Creation, viewed 01 March 2016, from https://github.com/marytts/marytts/wiki/HMMVoiceCreation.

Google maps, viewed 13 June 2016 from https://www.google.co.za/maps/@-23.8381146,28.6868749,8z.

Hassan, H. S., & Manap, F. A., 2014, ‘Listening to German Native and Non-Native Speakers: An Evaluation of Students’ Comprehension’, *International Conference on Knowledge-Innovation-Excellence: Synergy in Language Research and Practice. (SoLLS.INTEC.13). Proceeding – Social and Behavioral Sciences 118 2014*, 159 – 165.

Huang, X., Acero, A., & Hon, H., 2001, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Jane Bonnell, Ed. New Jersey: Prentice Hall PTR.

Human Language Technology Research Group, Council for Scientific and Industrial Research (CSIR) Meraka Institute, *Qfrency*, viewed 05 April 2017, from http://www.qfrency.com/demo/index.php.

Janska, A.C., & Clark, R.A.J., 2010, 'Native and Non-native Speaker Judgements on the Quality of Synthesized Speech', *Proceedings of International Symposium of Computer Architecture (ISCA)*, 26-30 September 2010. Makuhari, Chiba, Japan. 1121 – 1124.

Kominek, J. & Black A.W., 2003, 'The CMU Artic speech databases for speech synthesis research', Tech. Rep. CMU-LTI-03-177, http://festvox.org/cmu_arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

Kominek, J. & Black, A.W., 2004, 'The CMU Arctic speech databases', *Proceedings of the 5th ISCA Speech Synthesis Workshok-2014*, Pittsburgh, PA, USA, 223 – 225.

Kröger, B.J., & Birkholz, P., 2009, 'Articulatory Synthesis of Speech and Singing: State of the Art and Suggestions for Future Research', *Multimodal Signals, LNAI5398*, 2009, 306 – 319.

Laura, 2015, Voice import tools tutorial: How to build a new voice with voice import tools, viewed 30 January 2016, from https://github.com/marytts/marytts/wiki/VoiceImportToolsTutorial.

Lehohla, P., 2012, Statistics South Africa, Census 2011 Census in brief, report no.: 03-01-41, viewed 25 June 2017 from http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf.

Lemmetty, S., 1990, 'Review of speech synthesis technology', Department of Electrical and Communications Engineering, Helsinki University of Technology, viewed 19 April 2017, from http://research.spa.aalto.fi/publications/theses/lemmetty_mst/thesis.pdf.

Le Maguer, S,. Steiner, I., 2017, Hewer,A., ‘An HMM/DNN Comparison for Synchronized Text-to-speech and Tongue Motion Synthesis’, *Proceedings of Interspeech 2017*, August 20–24, 2017, Stockholm, Sweden, 239 – 243.

Louw, J.A., Davel, M., & Barnard E., 2005, ‘A general-purpose isiZulu speech synthesiser’, *South African Journal of African Languages*, 25 (2), 2015/1/1, Taylor & Francis Group, 92-100.

Madiba, M.R., 1994, ‘*A Linguistic Survey of Adoptives in Venda*’, Unpublished Masters Dissertation submitted at University of South Africa, viewed 03 March 2017, from http://uir.unisa.ac.za/bitstream/handle/10500/17284/dissertation_madiba_mr.pdf?sequence=1&isAllowed=y.

Mazzoni, D., 1999, *Audacity*, viewed 05 February 2016, from http://www.audacityteam.org/download.

McAvinue, S., 2014, ‘NZ voice wanted in a text to speech’, Otago Daily Times, 24 April, viewed 03 March 2016, from http://www.odt.co.nz/news/dunedin/299714/nz-voice-wanted-text-speech.

Morizane, K., Nakamura, K., Toda, T., Saruwatari, H. & Shikano, K., 2009, ‘Emphasized Speech Synthesis Based on Hidden Markov Models’, *Proceedings of Oriental COCOSDA 2009*, 76–81.

National African Language Resource Center, 2015, viewed 07 February 2016, from http://www.nalrc.indiana.edu/brochures/new-broch/Sepedi.pdf.

Oshima, Y, Takamichi, S, Toda, T, Neubig, G, Sakti, S & Nakamura, S., 2015, ‘Non-native speech synthesis preserving speaker individual based on partial correction of prosodic and phonetic characteristics’, *Proceedings of Interspeech Sep 2015*, Dresden, Germany, 299 – 303. Available: www.phontron.com/paper/oshima15interspeech.pdf.

Oura, K., 2006, *HMM-base speech synthesis system (HTS)*, HDecode-3.4.1.tar.gz, viewed 30 January 2016, from http://hts.sp.nitech.ac.jp/ftp/software/hdecode/HDecode-3.4.1.tar.gz.

Oura, K., 2006, *HMM-base speech synthesis system (HTS)*, HTS-2.2_for_HTK-3.4.1.tar.bz2, viewed 30 January 2016, from http://hts.sp.nitech.ac.jp/archives/2.2/HTS-2.2_for_HTK-3.4.1.tar.bz2.

Pammi, S., Charfuelan, M. & Schröder, M., 2010, 'Multilingual Voice Creation Toolkit for the MARY TTS Platform', *Proceedings of International Conference in Language Resources and Evaluation*, Malta, 2010.

Qian, Y., Soong F.K., 2014, 'Tutorial 4: Deep Learning for Speech Generation and Synthesis', The 9th International Symposium on Chinese Spoken Language Processing, September 12-14, 2014, Singapore, viewed 10 October 2018, from https://www.superlectures.com/iscslp2014/tutorial-4-deep-learning-for-speech-generation-and-synthesis

ReadSpeaker Holding B.V., updated 28 March 2017, *Try Some of Our Text to Speech Voices*, (voice demo), viewed 28 November 2015, fromhttp://www.readspeaker.com/voice-demo/.

Resource Management Agency (RMA), 2013, Lwazi English tts corpus, viewed 01 February 2016, from http://rma.nwu.ac.za/index.php/resource-catalogue/lwazi2-eng-tts-corpus.html.

Sasirekha, D. & Chandra, E., 2012, 'Text to Speech: A Simple Tutorial', *International Journal of Soft Computing and Engineering (IJSCE)*, (2) 1, 275 – 278.

Schröder, M. & Trouvain, J., 2001, 'The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching', viewed 01 February 2017, from http://lnv-90208.sb.dfki.de/documentation/publications/schroeder_trouvain2001.pdf

Slashdot Media, 2017, *Sourceforge: Find, Create, and Publish Open Source software for free***,** hts_engine_API-1.05.tar.gz, viewed 30 January 2016, from http://sourceforge.net/projects/hts-engine/files/hts_engine_API-1.10/hts_engine_API-1.05.tar.gz/download.

Slashdot Media, 2017, *Sourceforge: Find, Create, and Publish Open Source software for free*, sptk/SPTK-3.4.1.tar.gz, viewed 30 January 2016, from http://downloads.sourceforge.net/sptk/SPTK-3.4.1.tar.gz.

Speech Processing, Speech Recognition Acoustic modeling Pronunciation dictionary, viewed 24 June 2016 from http://www.speech.cs.cmu.edu/15-492/slides/06_asr_am.pdf.

Stan, A., Yamagishi, J., King, S. & Aylett, M., 2010, 'The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate', *In Speech Communication, Elsevier 2011*, 53 (3), 442 – 450.

Steiner, I., Le Maguer, S., Manzoni, J., Gilles, P. & Trouvain, J., 2017, 'Developing new language tools for MARYTTS: The case of Luxembourgish', *Conference Electronic Voice Processing*, 2017, 186 – 192.

The Apache software foundation, (2017), *Apache Maven Project*, viewed 05 February 2016, from https://maven.apache.org/.

The University of Edinburgh The Centre for Speech Technology Research, *The Festival Speech Synthesis System*, viewed 18 May 2016, from www.cstr.ed.ac.uk/projects/festival/.

The University of Edinburgh: The Centre for Speech Technology Research, *Festival text-to-speech online demo – technical*, viewed 17 August 2015, from http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html.

Thomas, S., 2007, '*Natural sounding text-to-speech synthesis based on syllable-like units*', Master of science thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, viewed 19 April 2017, from https://www.clsp.jhu.edu/~samuel/pdfs/ms-thesis.pdf.

Thomas, S., Murthy, H. A., & Sekhar, C. C., 2005, 'Distribution Text to Speech Synthesis for Embedded Systems – An analysis', *Proceedings of the Eleventh National Conference on Communication: NCC-2005*, 273 – 276.

Tomokiyo, L. M., Black, A. W. & Lenzo, K. A., 2005, ‘Foreign accents in synthetic speech: development and evaluation’, *Proceedings of Interspeech 2005*, Lisbon, Portugal, 1469-1472.

Traunmüller, H., 2000, *Wolfgang von Kempelen’s speaking machine and its successors*, Viewed 29 February 2016, from http://www2.ling.su.se/staff/hartmut/kemplne.htm.

United Nations, 2016, Official Languages, viewed 18 November 2016, from http://www.un.org/en/sections/about-un/official-languages/.

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K., 2016, ‘WaveNet: A Generative Model for Raw Audio’, viewed 22 March 2017, from https://deepmind.com/blog/wavenet-generative-model-raw-audio/.

Walter, D., 2016, *Greenbot: How to get started with Google Text-to-speech*, IDG Communications, Inc., 2017, viewed 11 April 2017, from http://www.greenbot.com/article/2105862/android/how-to-get-started-with-google-text-to-speech.html.

Watson, C., Liu, W. & MacDonald, B., 2013, ‘The effect of age and native speaker status on synthetic speech intelligibility’, *Proceedings of the 8th ISCA Speech Synthesis Workshop 2013*, Barcelona, Spain, 195 – 200.

Weinberger, S.H., updated: 28 April 2017, *The Speech Accent Archive*, (George Mason University), (10 August 2015), viewed 17 August 2015, from http://accent.gmu.edu/browse_language.php?function=find&language=english.

Yamagishi, J., 2006, ‘An Introduction to HMM-Based Speech Synthesis’, viewed 10 March 2017, from https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/TrajectoryModelling/HTS-Introduction.pdf.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., & Tokuda, K., 2007, ‘The HMM-based speech synthesis system (HTS) version 2.0’, *Proceedings of the 6th ISCA Speech Synthesis Workshop (SSW-6)* August 2007, 294 – 299.

Zerbian, S., 2007, ‘A First Approach to Information Structure in Xitsonga/Xichangana’. *SOAS Working Papers in Logistics Vol.15*. 65 – 78.

# APPENDICES

## APPENDIX A – VOICE PREPARATION, AUTOLABELING AND TRAINING.

1. Place all the audio files from cmu_us_slt_arctic/wav to voice/data/wav into /voice/data/wav and the text from cmu_us_slt_arctic/etc/txt.done.data into /voice/data/txt.done.data
2. The server.sh script starts MaryTTS server
3. Run the import.sh script to open the Database import window.

*The following steps requires interaction with the interface:*

4. Choose the directory /voice/data
5. Set parameters

db.marybase: /voice/source/marytts/

db.estDir: /voice/source/speech_tools/

db.samplingrate: 16000

db.locale: en_US

db.gender: male

db.domain: general

db.maryServerHost: localhost

db.maryServerPort: 59125

db.voicename: slt

HMMVoiceCompiler.mavenBin /voice/soft/maven/bin/mvn

EHMMLabeler.ehmmDir /voice/source/marytts/lib/external/ehmm

6. Select and run “PraatPitchmarker”.

*For a male voice set the minimum pitch to 75 and maximum pitch to 300 and for a female voice set the minimum pitch to 100 and the maximum pitch to 500.*

Results: voice/pm/* files

7. Select and run “MCEPMaker”

*Results: voice/mcep/*.mcep files*

8. Select and run “Festvox2MaryTranscripts

**Voice compiling: Autolabeling**

9. Select and run “AllophonesExtractor”.

*Results: voice/prompt_allophones/*.xml.*

10. Change the path to /voice/source/marytts/lib/external/ehmm then

select and run “EHMMLabeler”

*Results: voice/ehmm/* files*

11. Set the threshold to 10 then
select and run “LabelPauseDeleter”.

*Results: voice/lab/*.lab*

12. Select and run “PhoneUnitLabelComputer”.

*Results: voice/phonelab/*.lab*

13. Select and run “TranscriptionAligner”.

*Results: voice/allphones/*.xml*

14. Select and run “FeatureSelection”.

*Results: voice/mary/features.txt*

15. Select and run “PhoneUnitFeatureComputer”.

*Results: voice/phonefeatures/*.pfeats*

16. Select and run “PhoneLabelFeatureAligner”.

*Results:*

> *phonefeatures directory*
> *phonelab directory*
> *mary/features.txt file*
> *$marytts/lib/external/externalBinaries.config*

**Voice compiling: voice training**

17. Select and run “HMMVoiceDataPrepation”.
18. Select and run “HMMVoiceConfigure”.
19. Select and run “HMMVoiceFeatureSelection”.

*Results: mary/hmmfeatures.txt*

20. Select and run “HMMVoiceMakeData”.
21. Select and run “HMMVoiceMakeVoice”.
22. Change the path to /voice/soft/maven/bin/mvn the
Select and run “HMMVoiceCompiler”.

We copy the zip and xml files using copying.sh

To install the developed synthetic voice we run installer.sh.

## APPENDIX B – SOUTH AFRICAN ENGLISH LOCALE ADDITION.

```
Locale = en_US en_GB en_ZA\
Module.classes.list = \
marytts.language.en.JTokeniser \
marytts.language.en.Preprocess \
marytts.module.en.JPhonemiser(en_US) \
marytts. module.en.JPhonemiser(en_GB) \
marytts. module.en.JPhonemiser(en_ZA) \
marytts.language.en.Prosody \
marytts. module.SimplePhoneme2AP(en_US) \
marytts.language.en.PronunciationModel \
marytts. module.OpenNLPPosTagger(en,en.pos) \
```

In module settings we set:

```
en_ZA.allophoneset = jar:/marytts/language/en_ZA/lexicon/ \
allophones.en_ZA.xml
en_ZA.userdict = MARY_BASE/user-dictionaries/userdict-en_ZA.txt
en_ZA.lexicon = jar:/marytts/language/en_ZA/lexicon/en_ZA_lexicon.fst
en_ZA.lettertosound = jar:/marytts/language/en_ZA/lexicon/en_ZA.lts
Featuremanager.classes.list = \
Marytts.features.FeatureProcessorManager(en_US) \
Marytts.features.FeatureProcessorManager(en_GB) \
Marytts.features.FeatureProcessorManager(en_ZA)
```

To open the transcription tool we execute the transcriptiontool.sh, then open the files languagefile.sh to edit them. To install the created English locale we execute languageinstallation.sh. The import.sh is also used to incorporate the newly developed ZA English locale in the MARY TTS system.

## APPENDIX C – TRAINING TEXT CORPUS

( arctic_a0001 "Author of the danger trail, Philip Steels, etc." )

( arctic_a0002 "Not at this particular case, Tom, apologized Whittemore." )

( arctic_a0003 "For the twentieth time that evening the two men shook hands." )

( arctic_a0004 "Lord, but I'm glad to see you again, Phil." )

( arctic_a0005 "Will we ever forget it." )

( arctic_a0006 "God bless 'em, I hope I'll go on seeing them forever." )

( arctic_a0007 "And you always want to see it in the superlative degree." )

( arctic_a0008 "Gad, your letter came just in time." )

( arctic_a0009 "He turned sharply, and faced Gregson across the table." )

( arctic_a0010 "I'm playing a single hand in what looks like a losing game." )

( arctic_a0011 "If I ever needed a fighter in my life I need one now." )

( arctic_a0012 "Gregson shoved back his chair and rose to his feet." )

( arctic_a0013 "He was a head shorter than his companion, of almost delicate physique." )

……………

( arctic_a0590 "In a way he is my protege." )

( arctic_a0591 "We are both children together." )

( arctic_a0592 "It's only his indigestion I find fault with." )

( arctic_a0593 "She'd make a good wife for the cashier." )

( arctic_b0001 "Gad, do I remember it." )

( arctic_b0002 "You got out by fighting, and I through a pretty girl." )

( arctic_b0003 "I can see that knife now." )

( arctic_b0004 "When I can't see beauty in woman I want to die." )

( arctic_b0005 "His slim fingers closed like steel about Philip's." )

( arctic_b0006 "He seized Gregson by the arm and led him to the door." )

( arctic_b0007 "Hear the Indian dogs wailing down at Churchill." )

( arctic_b0008 "Burke himself had criticized it because of the smile." )

……………

( arctic_b0531 "I am sure it must have been some adventure." )

( arctic_b0532 "That Longfellow chap most likely had written countless books of poetry." )

( arctic_b0533 "His abnormal power of vision made abstractions take on concrete form." )

( arctic_b0534 "I'll tell you, the librarian said with a brightening face." )

( arctic_b0535 "He read his fragments aloud." )

( arctic_b0536 "Typhoid -- did I tell you." )

( arctic_b0537 "But she had become an automaton." )

( arctic_b0538 "At the best, they were necessary accessories." )

( arctic_b0539 "You were making them talk shop, Ruth charged him." )

## APPENDIX D – ALLOPHONES_EN_ZA.XML

```
<allophones name="sampa" xml:lang="en-ZA"
            features="vlng vheight vfront vrnd ctype cplace cvox">

      <!--
            VOWEL
                  length = vlng
                  vowel height = vheight
                  vowel frontness = vfront
                  vowel lip rounding = vrnd
                  vowel stress = stressed
            CONSONANT
                  type = ctype
                  place of articulation = cplace
                  voicing = cvox
                  aspiration = casp
            USAGE
                  vheight:[1,4]; vfront:[1,3]; vrnd:(-,+); stressed:(-,+)

                  vfront: 0=n/a  1=front      2=mid 3=back
                  vheight:0=n/a       1=high      2=mid-high 3=mid-low
                                                            4=low
                  vlng:   0=n/a  s=short      l=long d=diphthong a=schwa
                  vrnd:   0=n/a  +=on   -=off

                  ctype:(s, f, a, n, l, r) - consonant type: stop fricative affricative
                                              nasal liquid approximant
                  cplace:(l, a, p, b, d, v) - place of articulation: bilabial alveolar
                                              palatal labio-dental dental velar
                  cvox:(-,+); casp:(-,+); long:(-,+)
      -->

      <silence ph="_"/>
```

```
<vowel ph="A" vlng="l" vheight="3" vfront="3" vrnd="-"/>
<vowel ph="O" vlng="l" vheight="3" vfront="3" vrnd="+"/>
<vowel ph="u" vlng="l" vheight="1" vfront="3" vrnd="+"/>
<vowel ph="i" vlng="l" vheight="1" vfront="1" vrnd="-"/>
<vowel ph="i:" vlng="l" vheight="1" vfront="1" vrnd="-"/>
<vowel ph="u:" vlng="l" vheight="1" vfront="3" vrnd="+"/>

<vowel ph="{" vlng="s" vheight="3" vfront="1" vrnd="-"/>
<vowel ph="V" vlng="s" vheight="2" vfront="2" vrnd="-"/>
<vowel ph="E" vlng="s" vheight="2" vfront="1" vrnd="-"/>
<vowel ph="I" vlng="s" vheight="1" vfront="1" vrnd="-"/>
<vowel ph="U" vlng="s" vheight="1" vfront="3" vrnd="+"/>
<vowel ph="3:" vlng="s" vheight="3" vfront="1" vrnd="+"/>
<vowel ph="O:" vlng="s" vheight="3" vfront="3" vrnd="+"/>
<vowel ph="a" vlng="0" vheight="4" vfront="1" vrnd="-"/>
<vowel ph="A:" vlng="s" vheight="4" vfront="3" vrnd="-"/>
<vowel ph="Q" vlng="0" vheight="4" vfront="3" vrnd="+"/>


<vowel ph="@" vlng="a" vheight="2" vfront="2" vrnd="-"/>
<vowel ph="r=" vlng="a" vheight="2" vfront="2" vrnd="-" ctype="r"/>

<vowel ph="aU" vlng="d" vheight="3" vfront="2" vrnd="-"/>
<vowel ph="OI" vlng="d" vheight="2" vfront="3" vrnd="+"/>
<vowel ph="@U" vlng="d" vheight="2" vfront="3" vrnd="+"/>
<vowel ph="EI" vlng="d" vheight="2" vfront="1" vrnd="-"/>
<vowel ph="AI" vlng="d" vheight="3" vfront="2" vrnd="-"/>

<!-- Diphthongs consonants -->
<vowel ph="@i" vlng="d" vheight="2" vfront="2" vrnd="+"/>
<vowel ph="ai" vlng="d" vheight="2" vfront="2" vrnd="+"/>
<vowel ph="Oi" vlng="d" vheight="2" vfront="2" vrnd="+"/>
<vowel ph="@u" vlng="d" vheight="2" vfront="2" vrnd="+"/>
```

```
<vowel ph="au" vlng="d" vheight="2" vfront="2" vrnd="+"/>
<vowel ph="i@" vlng="d" vheight="2" vfront="2" vrnd="+"/>
<vowel ph="e@" vlng="d" vheight="2" vfront="2" vrnd="+"/>
<vowel ph="u@" vlng="d" vheight="2" vfront="2" vrnd="+"/>

<!-- Stop consonants -->
<consonant ph="p" ctype="s" cplace="l" cvox="-"/>
<consonant ph="t" ctype="s" cplace="a" cvox="-"/>
<consonant ph="k" ctype="s" cplace="v" cvox="-"/>
<consonant ph="b" ctype="s" cplace="l" cvox="+"/>
<consonant ph="d" ctype="s" cplace="a" cvox="+"/>
<consonant ph="g" ctype="s" cplace="v" cvox="+"/>

<!-- Affricates -->
<consonant ph="tS" ctype="a" cplace="p" cvox="-"/>
<consonant ph="dZ" ctype="a" cplace="p" cvox="+"/>
<consonant ph="d_0Z" ctype="a" cplace="a" cvox="+"/>

<!-- Affricative consonants -->
<consonant ph="f" ctype="f" cplace="b" cvox="-"/>
<consonant ph="v" ctype="f" cplace="b" cvox="+"/>
<consonant ph="T" ctype="f" cplace="d" cvox="-"/>
<consonant ph="D" ctype="f" cplace="d" cvox="+"/>
<consonant ph="s" ctype="f" cplace="a" cvox="-"/>
<consonant ph="z" ctype="f" cplace="a" cvox="+"/>
<consonant ph="S" ctype="f" cplace="p" cvox="-"/>
<consonant ph="Z" ctype="f" cplace="p" cvox="+"/>
<consonant ph="x" ctype="f" cplace="v" cvox="-"/>
<consonant ph="h" ctype="f" cplace="g" cvox="-"/>
<consonant ph="h\" ctype="f" cplace="g"/>

<!-- <consonant ph="l" ctype="l" cplace="a" cvox="+"/> -->

<!-- Nasal consonants -->
```

```
        <consonant ph="m" ctype="n" cplace="l" cvox="+"/>
        <consonant ph="n" ctype="n" cplace="a" cvox="+"/>
        <consonant ph="N" ctype="n" cplace="v" cvox="+"/>

        <!-- Approximant consonants (semivowels) -->
        <consonant ph="r" ctype="r" cplace="a" cvox="+"/>
        <consonant ph="r\" ctype="r" cplace="a"/>
        <consonant ph="w" ctype="r" cplace="l" cvox="+"/>
        <consonant ph="j" ctype="r" cplace="p" cvox="+"/>
        <consonant ph="l" ctype="r" cplace="a"/>
</allophones>
```

## APPENDIX E – VOICES REPORTS

The report obtained from praat using the same sentence for each synthetic and speaker voice.

**-- Voice report for Sound *xitsonga_female* –**

| | | Properties | |
|---|---|---|---|
| | | **Recording** | **Synthesised speech** |
| **Pitch** | Median pitch | 221.552 Hz | 221.253 Hz |
| | Mean pitch | 222.324 Hz | 221.979 Hz |
| | Standard deviation | 36.011 Hz | 33.410 Hz |
| | Minimum pitch | 132.798 Hz | 134.199 Hz |
| | Maximum pitch | 316.184 Hz | 312.519 Hz |
| **Pulses** | Number of pulses | 572 | 535 |
| | Number of periods | 551 | 524 |
| | Mean period | 4.511776E-3 seconds | 4.510488E-3 seconds |
| | Standard deviation of period | 0.759829E-3 seconds | 0.699112E-3 seconds |
| **Voicing** | Fraction of locally unvoiced frames | 31.156% (124 / 398) | 34.247% (125 / 365) |
| | Number of voice breaks | 11 | 9 |
| | Degree of voice breaks | 15.759% | 17.966% |

|  |  |  |  |
|---|---|---|---|
| **Jitter** | Jitter (local) | 2.288% | 1.144% |
|  | Jitter (local, absolute) | 103.223E-6 seconds | 51.611E-6 seconds |
|  | Jitter (rap) | 1.119% | 0.477% |
|  | Jitter (ppq5) | 1.399% | 0.511% |
|  | Jitter (ddp) | 3.356% | 1.432% |
| **Shimmer** | Shimmer (local) | 11.893% | 8.876% |
|  | Shimmer (local, dB) | 1.197 dB | 0.924 Db |
|  | Shimmer (apq3) | 4.864% | 2.364% |
|  | Shimmer (apq5) | 7.747% | 3.603% |
|  | Shimmer (apq11) | 13.710% | 8.498% |
|  | Shimmer (dda) | 14.593% | 7.091% |
| **Harmonicity of the voiced parts only** | Mean autocorrelation | 0.851447 | 0.941534 |
|  | Mean noise-to-harmonics ratio | 0.220574 | 0.071592 |
|  | Mean harmonics-to-noise ratio | 10.366 dB | 14.923 dB |

**-- Voice report for Sound *xitsonga_male* --**

|  | **Properties** |  |
|---|---|---|
|  | **Recording** | **Synthesised speech** |

| | | | |
|---|---|---|---|
| **Pitch** | Median pitch | 193.831 Hz | 190.945 Hz |
| | Mean pitch | 195.338 Hz | 183.198 Hz |
| | Standard deviation | 39.894 Hz | 31.003 Hz |
| | Minimum pitch | 107.984 Hz | 88.889 Hz |
| | Maximum pitch | 301.709 Hz | 240.981 Hz |
| **Pulses** | Number of pulses | 433 | 386 |
| | Number of periods | 427 | 380 |
| | Mean period | 5.112443E-3 seconds | 5.450929E-3 seconds |
| | Standard deviation of period | 1.096898E-3 seconds | 1.038563E-3 seconds |
| **Voicing** | Fraction of locally unvoiced frames | 26.923% (84 / 312) | 28.716% (85 / 296) |
| | Number of voice breaks | 4 | 5 |
| | Degree of voice breaks | 4.911% | 17.126% |
| **Jitter** | Jitter (local) | 2.840% | 1.620% |
| | Jitter (local, absolute) | 145.208E-6 seconds | 88.289E-6 seconds |
| | Jitter (rap) | 1.466% | 0.618% |
| | Jitter (ppq5) | 1.655% | 0.742% |

| | | | |
|---|---|---|---|
| | Jitter (ddp) | 4.397% | 1.853% |
| **Shimmer** | Shimmer (local) | 17.132% | 11.283% |
| | Shimmer (local, dB) | 1.488 dB | 1.186 dB |
| | Shimmer (apq3) | 7.691% | 2.612% |
| | Shimmer (apq5) | 12.029% | 4.135% |
| | Shimmer (apq11) | 18.945% | 13.119% |
| | Shimmer (dda) | 23.072% | 7.835% |
| **Harmonicity of the voiced parts only** | Mean autocorrelation | 0.850365 | 0.930474 |
| | Mean noise-to-harmonics ratio | 0.212122 | 0.089634 |
| | Mean harmonics-to-noise ratio | 9.211 dB | 14.452 dB |

**-- Voice report for Sound *tshivenda_female* –**

| | | **Properties** | |
|---|---|---|---|
| | | **Recording** | **Synthesised speech** |
| **Pitch** | Median pitch | 225.730 Hz | 220.017 Hz |
| | Mean pitch | 230.263 Hz | 226.295 Hz |
| | Standard deviation | 23.349 Hz | 31.320 Hz |
| | Minimum pitch | 189.947 Hz | 148.322 Hz |

| | | | |
|---|---|---|---|
| | Maximum pitch | 287.119 Hz | 295.803 Hz |
| **Pulses** | Number of pulses | 356 | 243 |
| | Number of periods | 351 | 234 |
| | Mean period | 4.336063E-3 seconds | 4.419754E-3 seconds |
| | Standard deviation of period | 0.460918E-3 seconds | 0.610383E-3 seconds |
| **Voicing** | Fraction of locally unvoiced frames | 55.241% (195 / 353) | 58.779% (154 / 262) |
| | Number of voice breaks | 4 | 7 |
| | Degree of voice breaks | 17.183% | 40.415% |
| **Jitter** | Jitter (local) | 1.996% | 0.799% |
| | Jitter (local, absolute) | 86.555E-6 seconds | 35.329E-6 seconds |
| | Jitter (rap) | 0.986% | 0.353% |
| | Jitter (ppq5) | 1.155% | 0.428% |
| | Jitter (ddp) | 2.957% | 1.058% |
| **Shimmer** | Shimmer (local) | 9.880% | 11.500% |
| | Shimmer (local, dB) | 1.109 dB | 0.957 dB |
| | Shimmer (apq3) | 3.692% | 2.693% |

|  |  |  |  |
|---|---|---|---|
|  | Shimmer (apq5) | 5.763% | 3.649% |
|  | Shimmer (apq11) | 10.498% | 12.399% |
|  | Shimmer (dda) | 11.077% | 8.078% |
| **Harmonicity of the voiced parts only** | Mean autocorrelation | 0.918529 | 0.960439 |
|  | Mean noise-to-harmonics ratio | 0.110332 | 0.049221 |
|  | Mean harmonics-to-noise ratio | 13.479 dB | 17.734 dB |

**-- Voice report for Sound *tshivenda_male* –**

|  |  | **Properties** |  |
|---|---|---|---|
|  |  | **Recording** | **Synthesised speech** |
| **Pitch** | Median pitch | 176.848 Hz | 171.526 Hz |
|  | Mean pitch | 173.686 Hz | 167.803 Hz |
|  | Standard deviation | 22.559 Hz | 25.954 Hz |
|  | Minimum pitch | 89.670 Hz | 122.962 Hz |
|  | Maximum pitch | 226.118 Hz | 239.192 Hz |
| **Pulses** | Number of pulses | 354 | 305 |
|  | Number of periods | 349 | 298 |

| | | | |
|---|---|---|---|
| | Mean period | 5.734163E-3 seconds | 5.967959E-3 seconds |
| | Standard deviation of period | 0.740921E-3 seconds | 0.945066E-3 seconds |
| **Voicing** | Fraction of locally unvoiced frames | 29.310% (85 / 290) | 28.063% (71 / 253) |
| | Number of voice breaks | 2 | 6 |
| | Degree of voice breaks | 2.071% | 12.869% |
| **Jitter** | Jitter (local) | 2.045% | 1.299% |
| | Jitter (local, absolute) | 117.273E-6 seconds | 77.531E-6 seconds |
| | Jitter (rap) | 0.964% | 0.523% |
| | Jitter (ppq5) | 1.270% | 0.558% |
| | Jitter (ddp) | 2.892% | 1.569% |
| **Shimmer** | Shimmer (local) | 13.361% | 11.240% |
| | Shimmer (local, dB) | 1.283 dB | 1.200 dB |
| | Shimmer (apq3) | 5.337% | 1.955% |
| | Shimmer (apq5) | 8.102% | 3.731% |
| | Shimmer (apq11) | 13.709% | 10.674% |
| | Shimmer (dda) | 16.010% | 5.866% |

| | | | |
|---|---|---|---|
| **Harmonicity of the voiced parts only** | Mean autocorrelation | 0.884148 | 0.952385 |
| | Mean noise-to-harmonics ratio | 0.170851 | 0.059791 |
| | Mean harmonics-to-noise ratio | 11.510 dB | 16.267 dB |

**-- Voice report for Sound *sepedi_female* --**

| | | **Properties** | |
|---|---|---|---|
| | | **Recording** | **Synthesised speech** |
| **Pitch** | Median pitch | 205.761 Hz | 183.138 Hz |
| | Mean pitch | 208.645 Hz | 193.393 Hz |
| | Standard deviation | 52.033 Hz | 32.209 Hz |
| | Minimum pitch | 80.207 Hz | 144.563 Hz |
| | Maximum pitch | 423.440 Hz | 286.233 Hz |
| **Pulses** | Number of pulses | 612 | 321 |
| | Number of periods | 579 | 310 |
| | Mean period | 4.889374E-3 seconds | 5.172980E-3 seconds |
| | Standard deviation of period | 1.295531E-3 seconds | 0.835159E-3 seconds |

| | | | |
|---|---|---|---|
| **Voicing** | Fraction of locally unvoiced frames | 25.339% (112 / 442) | 57.724% (213 / 369) |
| | Number of voice breaks | 12 | 10 |
| | Degree of voice breaks | 25.211% | 46.295% |
| **Jitter** | Jitter (local) | 2.752% | 0.986% |
| | Jitter (local, absolute) | 134.548E-6 seconds | 51.030E-6 seconds |
| | Jitter (rap) | 1.351% | 0.438% |
| | Jitter (ppq5) | 1.329% | 0.494% |
| | Jitter (ddp) | 4.052% | 1.315% |
| **Shimmer** | Shimmer (local) | 15.408% | 10.926% |
| | Shimmer (local, dB) | 1.383 dB | 0.739 dB |
| | Shimmer (apq3) | 5.471% | 3.130% |
| | Shimmer (apq5) | 9.466% | 4.520% |
| | Shimmer (apq11) | 23.909% | 6.511% |
| | Shimmer (dda) | 16.414% | 9.390% |
| **Harmonicity of the voiced parts only** | Mean autocorrelation | 0.856688 | 0.970104 |
| | Mean noise-to-harmonics ratio | 0.200499 | 0.032489 |
| | Mean harmonics-to-noise ratio | 9.669 dB | 18.218 dB |

**-- Voice report for Sound *sepedi_male* --**

| | | Properties | |
|---|---|---|---|
| | | **Recording** | **Synthesised speech** |
| **Pitch** | Median pitch | 162.648 Hz | 149.356 Hz |
| | Mean pitch | 160.246 Hz | 149.679 Hz |
| | Standard deviation | 14.582 Hz | 24.031 Hz |
| | Minimum pitch | 105.618 Hz | 104.397 Hz |
| | Maximum pitch | 195.356 Hz | 220.386 Hz |
| **Pulses** | Number of pulses | 346 | 297 |
| | Number of periods | 337 | 287 |
| | Mean period | 6.268711E-3 seconds | 6.675434E-3 seconds |
| | Standard deviation of period | 0.634205E-3 seconds | 1.042720E-3 seconds |
| **Voicing** | Fraction of locally unvoiced frames | 33.237% (115 / 346) | 42.735% (150 / 351) |
| | Number of voice breaks | 6 | 9 |
| | Degree of voice breaks | 19.056% | 35.455% |

|  |  |  |  |
|---|---|---|---|
| **Jitter** | Jitter (local) | 2.344% | 1.749% |
|  | Jitter (local, absolute) | 146.936E-6 seconds | 116.772E-6 seconds |
|  | Jitter (rap) | 1.104% | 0.731% |
|  | Jitter (ppq5) | 1.285% | 0.849% |
|  | Jitter (ddp) | 3.313% | 2.193% |
| **Shimmer** | Shimmer (local) | 14.496% | 10.834% |
|  | Shimmer (local, dB) | 1.367 dB | 1.130 dB |
|  | Shimmer (apq3) | 4.746% | 2.797% |
|  | Shimmer (apq5) | 9.043% | 4.430% |
|  | Shimmer (apq11) | 19.408% | 12.487% |
|  | Shimmer (dda) | 14.238% | 8.391% |
| **Harmonicity of the voiced parts only** | Mean autocorrelation | 0.884657 | 0.921050 |
|  | Mean noise-to-harmonics ratio | 0.155373 | 0.104991 |
|  | Mean harmonics-to-noise ratio | 11.068 dB | 13.260 dB |

# APPENDIX F – QUESTIONNAIRE

## South African English Text-to-speech Voice Rating Questionnaire

Please complete the following questions honestly. Your identity is confidential and will never be revealed at any case. The opinion you give on these synthetic voices will assist the researcher in reporting about the work done.

Details of Participant

Full Name: ________________________________

Age range:

| | | | | | | |
|---|---|---|---|---|---|---|
| < 18 | 18 - 25 | 26 – 35 | 36 - 45 | 46 - 55 | 56 - 65 | > 65 |

Gender:

| | |
|---|---|
| Male | Female |

Home Language: ______________

Occupation:

| | | |
|---|---|---|
| Learner | Undergraduate | Postgraduate/<br>Employee |

Dear Candidate: This questionnaire consists of 3 sections which should be completed after listening to each voice.

## SECTION 1: Naturalness

This section tests the naturalness of the synthetic voices, listen to the sentences uttered and thereafter rate each voice using the 5-point scale

Scale: **1**-completely unnatural, **2**-unnatural, **3**-average, **4**-natural, **5**-completely natural

1.1. How do the synthetic voices sound:

| | | | | | | |
|---|---|---|---|---|---|---|
| 1.1.1. | Voice *x_f* | 1 | 2 | 3 | 4 | 5 |
| 1.1.2. | Voice *x_m* | 1 | 2 | 3 | 4 | 5 |
| 1.1.3. | Voice *t_f* | 1 | 2 | 3 | 4 | 5 |
| 1.1.4. | Voice *t_m* | 1 | 2 | 3 | 4 | 5 |
| 1.1.5. | Voice *s_f* | 1 | 2 | 3 | 4 | 5 |
| 1.1.6. | Voice *s_m* | 1 | 2 | 3 | 4 | 5 |

**SECTION 2: Intelligibility**

This section tests the intelligibility of the synthetic voices. In this section you are expected to write the sentences each synthetic voice has spoken before you rate the voices.

2.1. Write down the sentence synthesised by each voice:

2.1.1. Voice *xitsonga_female*:

a. ________________________________________

b. ________________________________________

2.1.2. Voice *xitsonga_male*:

a. ________________________________________

b. ________________________________________

2.1.3. Voice *tshivenda_female*:

a. ________________________________________

b. ________________________________________

2.1.4. Voice *tshivenda_male*:

a. ________________________________________

b. ________________________________________

2.1.5. Voice *sepedi_female*:

a. ________________________________________

b. ________________________________________

2.1.6. Voice *sepedi_male:*

a. ________________________________________

b. ________________________________________

2.2. Based on question 2.1 above, rate each voice on a 5-point scale:

Scale: **1** – totally not understandable, **2** – not understandable, **3** – average, **4** – understandable, **5** – totally understandable

| | | | | | | |
|---|---|---|---|---|---|---|
| 2.2.1. | Voice x_f | 1 | 2 | 3 | 4 | 5 |
| 2.2.2. | Voice x_m | 1 | 2 | 3 | 4 | 5 |
| 2.2.3. | Voice t_f | 1 | 2 | 3 | 4 | 5 |
| 2.2.4. | Voice t_m | 1 | 2 | 3 | 4 | 5 |
| 2.2.5. | Voice s_f | 1 | 2 | 3 | 4 | 5 |
| 2.2.6. | Voice s_m | 1 | 2 | 3 | 4 | 5 |

**SECTION 3: Similarity**

This section is intended to check the similarity between the speaker and the synthetic voice created. In this section, an audio file from the recordings will be played and the same sentence will be synthesised. You are required to rate the level of similarity from the two files using the 5-point scale.

Scale: 1 – sounds like a total different speaker, 2 – sounds like a different speaker, 3 – sounds neutral, 4 – sounds like the speaker, 5 – sounds exactly like the speaker

3.1. The synthetic voice:

| | | | | | |
|---|---|---|---|---|---|
| 3.1.1. Voice x_f | 1 | 2 | 3 | 4 | 5 |
| 3.1.2. Voice x_m | 1 | 2 | 3 | 4 | 5 |
| 3.1.3. Voice t_f | 1 | 2 | 3 | 4 | 5 |
| 3.1.4. Voice t_m | 1 | 2 | 3 | 4 | 5 |
| 3.1.5. Voice s_f | 1 | 2 | 3 | 4 | 5 |
| 3.1.6. Voice s_m | 1 | 2 | 3 | 4 | 5 |

## APPENDIX G – CONFERENCE PAPERS

**Malatji, P.T.**, Manamela, M.J. & Sefara, T.J., 2016. Creating accented text-to-speech English voices to facilitate second language learning. In *South Africa International Conference on Educational Technologies*. Pretoria, AARF, pp. 234 – 242.

**Malatji, P.T.**, Manamela, M.J. & Sefara, T.J., 2017. Second language learning through accented synthetic voices. In *South Africa International Conference on Educational Technologies*. Pretoria, AARF, pp. 109 – 119.

Sefara, T.J., Manamela, M.J. & **Malatji, P.T.**, 2016. Applying Speech Synthesis to Basic Mathematics as a Language. In *South Africa International Conference on Educational Technologies*. Pretoria, AARF, pp. 243 – 253.

Sefara, T.J., Manamela, M.J. & **Malatji, P.T.**, 2016. Text-based language identification for some of the under-resourced languages of South Africa. In *3rd International Conference on Advances in Computing and Communication Engineering (ICACCE-2016)*. Durban, IEEE, pp. 308 – 312.

Sefara, T.J., Madimetja, J.M., Modipa, T.I., Mokgonyane, T.B., Manamela, P.J. & **Malatji, P.T.**, 2017. Speech synthesis systems for non-native language learning. In *2017 Africon*. Cape Town, (In press).