# HMM-based Speech Synthesis System incorporated with Language Identification for Low-resourced Languages

Tshephisho Joseph Sefara
*Modelling and Digital Science*
*Council for Scientific and Industrial Research*
Pretoria, South Africa
tsefara@csir.co.za

Tumisho Billson Mokgonyane
*Department of Computer Science*
*University of Limpopo*
Polokwane, South Africa
mokgonyanetb@gmail.com

Madimetja Jonas Manamela
*Department of Computer Science*
*University of Limpopo*
Polokwane, South Africa
jonas.manamela@ul.ac.za

Thipe Isaiah Modipa
*Department of Computer Science*
*University of Limpopo*
Polokwane, South Africa
thipe.modipa@ul.ac.za

***Abstract*—Text-to-speech (TTS) synthesis systems are of benefit towards learning new or foreign languages. These systems are currently available for various major languages but not available for low-resourced languages. Scarcity of these systems may lead to challenges in learning new languages specifically low-resourced languages. Development of language-specific systems like TTS and Language identification (LID) have an important task to address in mitigating the historical linguistic effects of discrimination and domination imposed onto low-resourced indigenous languages. This paper presents the development of a multi-language LID+TTS synthesis system that generate audio of input text using the predicted language in four South African languages, namely: Tshivenda, Sepedi, Xitsonga and IsiNdebele. On the front-end, is the LID module that detects language of the input text before the TTS synthesis module produces output audio. The LID module is trained on a 4 million words dataset resulted with 99% accuracy outperforming the state-of-the-art systems. A robust method for building TTS voices called hidden Markov model method is used to build new voices in the selected languages. The quality of the voices is measured using the mean opinion score and word error rate metrics that resulted with positive results on the understandability, naturalness, pleasantness, intelligibility and overall impression of the system of the newly created TTS voices. The system is available as a website service.**

***Keywords— text-to-speech synthesis, language identification, neural networks, machine learning, natural language processing***

## I. INTRODUCTION

There has been an improvement of speech processing technologies over the last few decades within the realm of human language technologies (HLTs), often called language technologies [1]. HLTs consist of computational linguistics or natural language processing on the front-end and speech technologies on the back-end. At the core of speech processing technologies lies the speech synthesis module, also referred to as text-to-speech (TTS) synthesis, and automatic speech recognition (ASR) module, also referred to as speech-to-text conversion. Rabiner and Juang [2] define ASR as a technology that recognises spoken language to produce equivalent text format. TTS synthesis is a process of generating spoken language by a computer system [3]. A computational system utilising this functionality is called a speech synthesiser, and can be implemented in a hardware or software product.

Most speech technologies are commercially available for well-resourced languages such as English and other European languages but such systems are limited for South African indigenous languages. About four official languages of South Africa namely, isiZulu, Sesotho, Afrikaans and isiXhosa are included in Google translate excluding the well-known global lingua franca, English. Although research in TTS synthesis systems in South Africa is relatively young and emerging, many TTS synthesis systems research efforts occurring in South Africa have acquired international awareness and exposure in terms of the quality and impact of the research work [4] [5] [6] [7].

These technologies enable machines to interact and communicate with humans, and deliver valuable and useful e-services ranging from sciences, health, economics, and education. The use of information and communication technology (ICT) is rapidly growing in educational institutions ranging from basic to tertiary education teaching and learning. Recent technologies encompass ICT and e-learning facilities delivered by modern tools including the internet, computers, smartphones, and Blackboard. These educational technologies enable students to improve their knowledge easily. Moreover, TTS synthesis functionality embedded in most mobile devices and computers can be used to learn additional languages. In a multilingual country like South Africa, additional language learning is still a challenge to most learners. Hence, the development of TTS synthesis systems covering all official South African languages may impact language learning and teaching. This paper mainly focuses on four South African languages spoken mainly in Limpopo province (Xitsonga, Sepedi, isiNdebele and Tshivenda) as these languages are regarded as low-resourced as stated in [8]. The contributions of this paper are as follows:

- Provide a platform [1] for new research in language learning and teaching of South African official languages.

[1] https://sefaratj.github.io/tts-lid

- Develop LID module that predicts language of input text under low-resourced environment.
- Develop a hidden Markov model (HMM)-based TTS synthesis system that is compatible with any web-based application.
- Evaluate the performance of the LID and quality of the synthetic voices using word error rate (WER) method in terms of intelligibility, and subjective mean opinion score (MOS) method in terms of pleasantness, naturalness, understandability, and overall impression of the system.

The rest of the paper is organised as follows. Section II discusses the literature review. Section III details the data and equipment used to implement the proposed system. Section IV details evaluation procedure. We discuss the results in Section V, and this paper is concluded in Section VI with recommendations.

## II. Literature Review

Speech technology applications can play an important role in teaching and language learning. An increase in the development of such system enhances learning through the use of computer-assisted language learning and computer-assisted pronunciation training (CAPT). Developing CAPT systems for pronunciation learning and teaching requires extensive linguistic resources and experts. Chen and Li [9] review approaches and challenges used in development of CAPT systems. A good analysis of using ASR for training students to learn new languages is discussed in [10]. Yu and Wang [11] propose a pronunciation visualisation instruction system based on articulatory mesh model. Their system is evaluated on students learning Chinese in second language and achieved accuracy of 97.6% (after learning) from 68.4% (before learning). This shows their system significantly impact learning of new languages.

In this digital age, computers are used in classrooms to facilitate teaching in all areas of language learning ranging from drawing, writing, and reading. Speech synthesis applications are being utilised in teaching English as a foreign language in classrooms. A system proposed by Malatji *et al*. [12] is the first speech-based system to focus on the effect of accent on English language learning in the South African context. Speech synthesis applications simplify language learning and pronunciation; such applications can be applied not only in language learning but extended to the domain of mathematics and science. An Android-based application was developed by Sefara *et al*. [13], their application synthesises mathematical equations into speech audio in Sepedi.

People with speech disorder use speech-based applications to conduct learning and daily communication. These applications are not available for all South African official languages. Hence, design, development and implementation of such applications are required in bridging the communication-divide. Dzulkifli *et al.* [14] review studies that use computer-assisted instructions (CAI) to enhance the language development of children with Autism Spectrum Disorder (ASD). Lack of vocabulary is a contributing factor for children being incompetent. Dzulkifli *et al*. [14] recommends using CAI on vocabulary learning of children with ASD. As HLT research improves and enhances methods of teaching and learning, new research initiatives at centres such as the Council for Scientific and Industrial Research (CSIR) enhances the development of summarisation, translation, TTS, ASR and LID systems for South African languages [15].

The use of LID module as a front-end service to a TTS synthesis system enhances and improves the performance when the origin language of the given input text is unknown. While most TTS synthesis systems such as IBM [2] and CereProc [3] do not have this service, Google translate [4] incorporates LID of input text using deep neural network (DNN) models. Different DNN-based implementations can solve the problem of language identification in text such as convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM), etc. Kim *et al*. [16] use character-level CNN to propose a light-weight neural language model with fewer parameters outperforming baseline models (like Kneser-Ney, word level/morpheme-level LSTM). With similar method, Zhang *et al*. [17] propose text classification using character-level CNNs.

## III. Methodology

This section details the data collections procedure, preparation of requirements, methods used to develop the synthetic voices and the LID module. Fig. 1 shows the overall system integration. The user inputs text on the user interface then the LID module predicts and sets the language that is used for speech generation. The TTS module generates synthetic speech based on the selected language.

### A. Data Collection and Acquisition

There are two requirements needed before building new synthetic voices. Firstly, data in a form of a pronunciation dictionary. Secondly, voice recordings or speech data together with their text annotations. These required secondary training speech data is obtained from the Lwazi (a name that means "knowledge" in isiZulu) project of language resource management agency[5] (RMA). The data for LID is also obtained from the same source. The LID data is textual data collected from the web. The summary of the LID data is shown in Table I with a total of **4 million words** after pre-processing the data as [18] shows that better performance can be obtained when the data is pre-processed by removing non-standard words, special characters, numbers, symbols, and Unicode characters left with alphabets only. The recruited and volunteering first language speakers were engaged for the recording of speech data. Table I show the number of transcriptions and recording duration respectively for selected four languages. The female recordings are in Sepedi and Tshivenda, while male recordings are in isiNdebele and Xitsonga. Total number of sentences in the corpora is 4222 sentences and the duration totals to 7 hours and 11 minutes.

### B. System Preparation

The Modular architecture for research on speech synthesis (MARY) TTS is one of the tools used to develop the synthetic voices. The MARY TTS synthesis system is a multi-language TTS synthesis system engine written in Java [19]. It is a collaborative project of the Saarland University

[2] https://text-to-speech-demo.ng.bluemix.net
[3] http://www.cereproc.com
[4] http://translate.google.com
[5] http://rma.nwu.ac.za

and German Research Centre for Artificial Intelligence (DFKI) [20]. This software is pre-installed with Deutsch, English, Italian, and other European languages. It has a support for the creation of new languages and building synthetic voices in unit selection and HMM-based method. We followed an HMM-based method for development of synthetic voices because of the following advantages; flexibility to change voice characteristics [21] [22], robustness [23] [24], small footprint [25] [26], and multilingual support [27] [28].

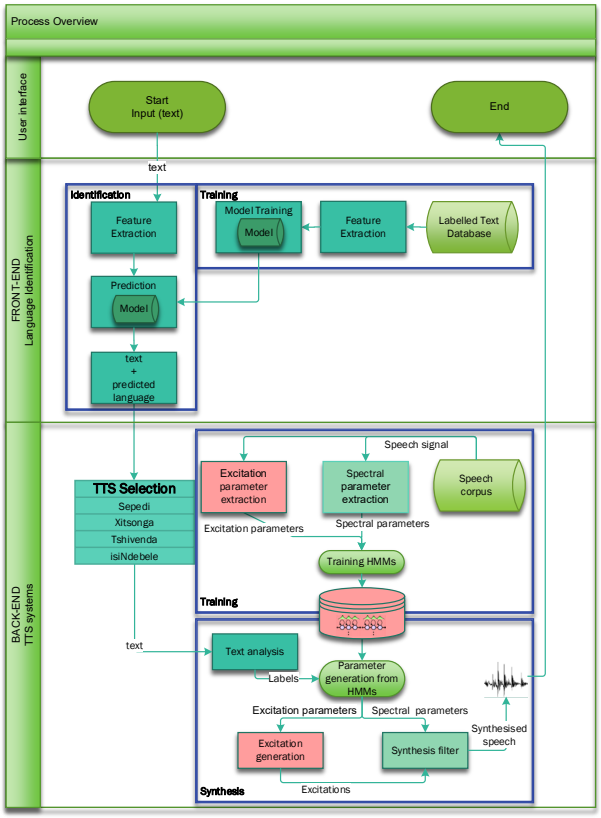


Fig. 1. System architecture showing user input text (surname) that passes through LID module for language detection. Then TTS selection module to sets a language for speech synthesis. Finally, the speech synthesis module generates synthetic speech.

TABLE I. SUMMARY OF CORPORA

| Language | LID Words | Transcriptions | Minutes |
|---|---|---|---|
| Sepedi | 1280595 | 1318 | 143 |
| Tshivenda | 1096897 | 1000 | 83 |
| IsiNdebele | 713121 | 994 | 117 |
| Xitsonga | 1201435 | 910 | 88 |
| **Total** | **4292048** | **4222** | **431** |

TABLE II. SAMPA PRONUNCIATION DICTIONARY

| Word | SAMPA | Type |
|---|---|---|
| tshepišo | ts_h E p_> I S O | |
| magetla | m a G E tl_> a | |
| wena | w E n a | functional |

The following tools are used for development of the proposed system: Ubuntu work station, Java development kit, Praat [29], HTK-3.4.1 and HDecode-3.4.1, HTS-2.2_for_HTK-3.4.1.patch, Edinburgh speech tools, SPTK, and HTS engine. The detailed instructions for installation of given tools are given in the file named INSTALL inside each software package, and also in the MARY TTS GitHub wiki page contains more installation instructions [30].

### C. Language Technology Module

The consonants and vowels (shown in Table II) for each language are used to create four phone set files following the Speech Assessment Methods Phonetic Alphabet (SAMPA) format. The phone sets contain SAMPA phones and their linguistic features. These features include length, height, frontness/backness, and mouth roundness for a vowel, and consonants have type and place of articulation, voicing, aspiration and palatal features. The pronunciation dictionary format is shown in Table II. The first word is followed by its SAMPA phonetic transcription then optionally ending with the word **functional**. The MARY TTS system contains a transcription tool that requires the pronunciation dictionary to create the following files (where *xy* is a language code under International Organisation for Standardisation ISO 639-2:2008) [30]: (a) *xy.lts* – Letter-to-Sound rule file used for transcription of unknown words. (b) *xy_lexicon.dict* – Phoneme-based pronunciation dictionary file. (c) *xy_lexicon.fst* – Grapheme-to-phoneme file uses rules to generate most probable phone list, sometimes shortened G2P. (d) *xy_pos.fst* – Part-of-Speech tagger used to classify and tag parts of sentences according to their classes including prepositions, articles, verbs, adjectives, nouns and others. Instructions given on New Language Support [30] from MARY TTS wiki page on GitHub are followed to implement the language components.

### D. Speech Coding and Implementation Platform

The speech corpus contains wave files recorded at a sampling rate of 16 kHz. This rate used mostly in voice communications over the internet for better performance. Additionally, the recordings are on a mono channel with 16 bit per sample. The transcription file contains sentences for all audio wave files and named correctly with the first word describing the audio file name followed by a sentence in double quotation describing the contents of the audio file. The MARY TTS synthesis system contains a voice import tool used to create new synthetic voices for the given languages. The tool requires the audio files and their transcription files. The voice training procedure may take more than 10 hours to finish creating a single HMM voice depending on the size of the speech corpus. Instructions given on Voice Import Tool Tutorial from MARY TTS GitHub wiki page were followed to create HMM-based voices [30].

### E. LID Model

We implement our models using TensorFlow (a deep learning toolkit) on the environment with 12GB memory and Tesla K80 GPU. We train two types of data mining models: (a) **Logistic Regression (LR)** the model contains a dense layer given few arguments; number of labels, and activation function as *softmax*. (b) **Deep Neural Network (DNN)** with input layer given the shape of the input features and a fully connected layer activated with *softmax* as the last layer shown in Table III. We train word2vec models for each language to help with data augmentation since the data is small. To show the importance of data augmentation, we augment 25% and 50% of the training data five times. This

increases the data up to 5×25% and 5×50% of the training data. We use TFIDF and one-hot encoding to prepare the input data.

TABLE III. TABLE OF THE DNN ARCHITECTURE

| Layer | Type | Filters/Neurons |
|---|---|---|
| 0 | Input | - |
| 1 | Fully connected+tanh | 512 |
| 2 | Dropout | - |
| 3 | Fully connected+tanh | 128 |
| 4 | Dropout | - |
| 5 | Fully connected+softmax | 4 |

## IV. EVALUATION

This section describes the procedure that is used for evaluation of the developed system.

### A. Mean Opinion Score

A 5-point Likert scale rating system is used to measure the quality of the synthetic voices, where 1 means "horrible" and 5 means "best". The evaluation questions in Fig. 2 are adopted from [31]. The MOS metric in (1) is used to measure understandability, pleasantness, naturalness and overall system impression. MOS is a performance metric applied to measure the quality of speech from subjective evaluations and the metric is defined as:

$$MOS = \frac{1}{n}\sum_{i=1}^{n} x \tag{1}$$

where $x$ is the score of the evaluator and $n$ is the total number of evaluators.

A sample of 32 native speakers evaluated the system in terms of understandability, pleasantness, naturalness, intelligibility and overall system impression. The subjects are university students and they are recruited via a word of mouth. The subjects are divided into four groups where each group is classified by their native language. The evaluations are divided into four sessions where each group of eight native speakers evaluate their native language TTS system. The subjects are given an opportunity to construct five standard sentences of their choice; and fed to the TTS system. They listen to the sentences once. Then subjects give their opinions according to the evaluation sheet in Fig. 2.

### B. Word Error Rate

The second evaluation is undertaken using the WER metric given in (2). WER is a common metric used to measure the performance of an automatic speech recognition and TTS synthesis systems on a word-level [3]. The WER metric is used in subjective evaluation of intelligibility of TTS synthesis systems. The metric is based on the minimum number of insertions, deletions and substitutions that have to be performed to convert the generated text (or hypothesis) into the reference text. The initial step in calculating word error is to find the minimum edit distance in words between the hypothesised and reference words [3]. Another five set of sentences are constructed using the semantically unpredictable sentences (SUS) and eight evaluators perform the evaluation for

**Evaluation Sheet**
**Date**
**Home Language**

| a) How much listening effort was needed to understand what was said? | b) How would you describe the pleasantness of the audio? | c) How would you rate naturalness of the audio? | d) How do you rate the quality of the overall system? |
|---|---|---|---|
| 1. Cannot understand<br>2. Maximum effort needed<br>3. Fair<br>4. Minimum effort needed<br>5. No effort needed | 1. Very unpleasant<br>2. Unpleasant<br>3. Satisfactory<br>4. Pleasant<br>5. Very pleasant | 1. Very unnatural<br>2. Unnatural<br>3. Satisfactory<br>4. Natural<br>5. Very natural | 1. Horrible<br>2. Poor<br>3. Tolerable<br>4. Good<br>5. Excellent |

Fig. 2. Evaluation Sheet

intelligibility of each synthetic speech. A total of 5 × 8 = 40 sentences per language are used during the evaluation. The total number of words are different per language. The total number of words for Sepedi, Xitsonga, Tshivenda, and isiNdebele were 312, 240, 248, and 176 respectively. The sentences were hidden from the subjects. Subjects listened to sentences and asked to type the sentences on a computer. The WER was applied to the sentences from SUS to measure intelligibility which is typically captured by the WER metric formulated as:

$$WER = \frac{S+I+D}{N} \times 100\% \tag{2}$$

where $S$ is the number of word substitution errors, $I$ is the number of word insertion errors, $D$ is the number of word deletion errors, and $N$ is the total number of words. The implementation of WER is a python script that receives a textual file containing hypothesis and reference sentences in a sequence. The WER and MOS results are averaged and Section V discusses the results.

### C. LID testing

The LID data is split into training and testing. The percentage split is 90% for training, 10% for testing. The training data is further split into 10k for evaluation and the rest for training. The following metrics are used to measure the performance of the model. (a) The **categorical cross-entropy** loss function using *softmax* activation function, and *adadelta* as the optimizer. (b) The evaluation **accuracy**.

## V. EXPERIMENTAL RESULTS AND DISCUSSIONS

### A. Results for TTS

The subjective MOS has reported good results for measuring the quality of speech over several factors. Synthetic speech is measured for understandability, pleasantness, naturalness, intelligibility, and overall system impression. Understandability describes how understandable is the synthetic voice. Understandability is sometimes called listening effort. Pleasantness focuses on pleasantness in the synthetic voice, while naturalness describes how natural is the synthetic voice compared to spoken language. Intelligibility focuses on the ability for people to understand the synthesised speech. The TTS synthesis systems trained with thousands of speech data are likely to produce synthetic voices that are rich in naturalness.

TABLE IV. MOS AND WER RESULTS FOR SEPEDI, ISINDEBELE, TSHIVENDA, AND XITSONGA.

| Language | Naturalness | Pleasantness | Understandability | Overall Quality | WER |
|---|---|---|---|---|---|
| Sepedi | 3.875 | 3.875 | 4.125 | 4.125 | 14.816 |
| IsiNdebele | **4.875**[a] | **4.75**[a] | **4.875**[a] | **4.75**[a] | **5.12**[a] |
| Tshivenda | 4.5 | 4.25 | 4.25 | 4.25 | 8.288 |
| Xitsonga | 4.625 | 4.625 | 4.375 | 4.625 | 5.914 |

[a.] Higher MOS and lower WER is better

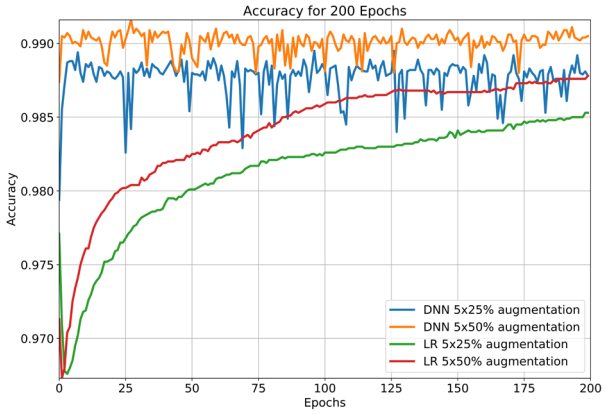


Fig. 3. Accuracy of the models

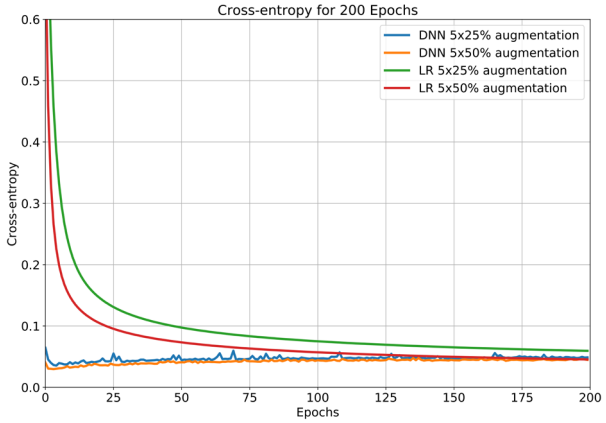


Fig. 4. Cross-entropy loss of the models

Table IV details the results for four languages after subjective evaluation. The MOS evaluation yield a mean of at least 4.13 for the four synthetic voices on understandability. Hence, this shows that the correct messages are conveyed. A mean pleasantness of at least 3.8 is scored for the four synthetic voices with isiNdebele scoring the mean of 4.75, showing that there are less robotic sounds in the isiNdebele synthetic speech compared to other languages. A mean naturalness of at least 4.25 is observed for Xitsonga, Tshivenda, and isiNdebele, showing that synthetic speech is more natural excluding Sepedi with the lowest mean of 3.8. Naturalness can be improved by using a professional speaker with a high-quality microphone and by using a handcrafted phone alignment during the creation of a synthetic voice. High-quality TTS synthesis systems are produced when creating new languages with correct pronunciation. The overall impression of the system scored a mean of above 4.13 with isiNdebele having higher mean of 4.75, according to MOS rating system results in Table IV showing that the system is acceptable and isiNdebele performed better than other languages with Xitsonga being the second better language followed by Tshivenda, and lastly, Sepedi.

The intelligibility of synthetic speech is measured using the WER metric. A good TTS synthesis system has the lowest WER. Table IV shows the results of the WER. Sepedi obtained higher WER of 14.82% compared to other languages. This may be caused by the familiarity of the speaker with the language since the speaker has background or accent of Setswana. Other languages obtained lower WER of below 9%. From these results, we may conclude that all the built synthetic voices are intelligible based on the subjective opinions.

### *B. Results for LID module*

We trained the model on 200 epochs or iterations using batch size of 1024. Fig. 3 shows the accuracy for 200 epochs for both LR and DNN. For LR, data augmentation plays important role to obtain better accuracy. We realise the accuracy after augmenting 50% training data is better than augmenting 25% training data. For DNN, we observe no improvement from data augmentation. This occurs when the DNN already seen features of the training data. Hence, more features may result in overfitting and noise. Fig. 4 shows the validation cross-entropy loss function against the epochs. We observe both LR and DNN smoothly decaying. This validation cross-entropy loss shows that the models were not overfitted. On the final test data, LR obtained accuracy of 99% and cross-entropy loss of 0.05, while DNN obtained same accuracy of 99% and better cross-entropy loss of 0.04. With these results, the DNN model is incorporated with the TTS for online demonstration.

## VI. CONCLUSION AND RECOMMENDATIONS

We have presented the development of TTS synthesis system for Xitsonga, Tshivenda, Sepedi and isiNdebele incorporating LID module on the front-end. We evaluated the system using tertiary students and received positive ratings. In conclusion, it can be said that with a newly created voice with high understandability, pleasantness, naturalness, intelligibility and generally acceptable evaluation results is attainable. The HMM-based TTS synthesis systems for indigenous South African languages can be used in real-life applications. The developed prototype system is among the first systems that contain multiple languages ever built for under-resourced languages of South Africa. The developed multi-language TTS synthesis system can be used as a platform in the ICT and e-learning institutions to enhance language learning and teaching. The future work will focus on adding more African languages and to apply these technologies in the teaching and learning institutions to measure their effectiveness. The online demo can be found on GitHub [32].

## ACKNOWLEDGMENT

This work is based on research supported by the University of Limpopo Centre of Excellence for Speech Technology and Council for Scientific and Industrial Research.

## REFERENCES


[1] P. Mukherjee, S. Santra, S. Bhowmick, A. Paul, P. Chatterjee and A. Deyasi, "Development of GUI for text-to-speech recognition using natural language processing," in *2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), IEEE*, Kolkata, India, 2018.

[2] L. R. Rabiner and B.-H. Juang, Fundamentals of speech recognition, vol. 14, PTR Prentice Hall Englewood Cliffs, 1993.

[3] D. Jurafsky and J. H. Martin, Speech and language processing, vol. 3, Pearson London, 2014.

[4] F. de Wet, W. van der Walt, N. Dlamini and A. Govender, "Building synthetic voices for under-resourced languages: The feasibility of using audiobook data," in *Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), IEEE*, Bloemfontein, South Africa, 2017.

[5] T. B. Mokgonyane, T. J. Sefara, P. J. Manamela, M. J. Manamela and T. I. Modipa, "Development of a speech-enabled basic arithmetic m-learning application for foundation phase learners," in *AFRICON, IEEE*, Cape Town, South Africa, 2017.

[6] P. J. Manamela, M. J. Manamela, T. I. Modipa, T. J. Sefara and T. B. Mokgonyane, "The automatic recognition of Sepedi speech emotions based on machine learning algorithms," in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), IEEE*, Durban, South Africa, 2018.

[7] T. J. Sefara, M. J. Manamela and P. T. Malatji, "Text-based language identification for some of the under-resourced languages of South Africa," in *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE), IEEE*, Durban, South Africa, 2016.

[8] L. Besacier, E. Barnard, A. Karpov and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication,* vol. 56, pp. 85-100, 2014.

[9] N. F. Chen and H. Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), IEEE*, Jeju, South Korea, 2016.

[10] M. Eskenazi, "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype," *Language Learning and Technology,* vol. 2, no. 2, pp. 62-76, 1999.

[11] J. Yu and Z. Wang, "A realistic and reliable 3D pronunciation visualization instruction system for computer-assisted language learning," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Shenzhen, China, 2016.

[12] P. T. Malatji, M. J. D. Manamela and T. J. Sefara, "Creating accented text-to-speech English voices to facilitate second language learning," in *South Africa International Conference on Educational Technologies*, Pretoria, South Africa, 2016.

[13] T. J. Sefara, M. J. Manamela and P. T. Malatji, "Applying speech systhesis to basic mathematics as a language," in *South Africa International Conference on Educational Technologies*, Pretoria, South Africa, 2016.

[14] M. A. Dzulkifli, E. V. F. Abdul and A. W. A. Rahman, "A review for future research and practice in using computer assisted instruction on vocabulary learning among children with autism spectrum disorder," in *2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M), IEEE*, Jakarta, Indonesia, 2016.

[15] E. Barnard, M. H. Davel and G. B. Van Huyssteen, "Speech technology for information access: a South African case study.," in *AAAI Spring Symposium: Artificial Intelligence for Development*, Palo Alto, California, 2010.

[16] Y. Kim, Y. Jernite, D. Sontag and A. M. Rush, "Character-aware neural language models," in *Proceeding AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, 2016.

[17] X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification," in *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* , Montreal, Canada, 2015.

[18] V. V. Nhlabano and P. E. Lutu, "Impact of text pre-processing on the performance of sentiment analysis models for social media data," in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), IEEE*, Durban, South Africa, 2018.

[19] S. Pammi, M. Charfuelan and M. Schröder, "Multilingual voice creation toolkit for the MARY TTS platform," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

[20] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology,* vol. 6, pp. 365-377, 2003.

[21] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems,* vol. 90, no. 2, pp. 533-543, 2007.

[22] M. Sulír and J. Juhár, "Speaker adaptation for Slovak statistical parametric speech synthesis based on hidden Markov models," in *2015 25th International Conference Radioelektronika (RADIOELEKTRONIKA), IEEE*, Pardubice, Czech Republic, 2015.

[23] J. Yamagishi, Z. Ling and S. King, "Robustness of HMM-based speech synthesis," in *INTERSPEECH-2008, 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, 2008.

[24] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, no. 1, pp. 66-83, 2009.

[25] H. Zen, A. Senior and A. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE*, Vancouver, BC, 2013.

[26] K. Hashimoto, K. Oura, Y. Nankaku and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[27] M. Gibson, T. Hirsimäki, R. Karhila, M. Kurimo and W. Byrne, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010.

[28] M. Gibson and W. Byrne, "Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 895-904, 2011.

[29] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International,* vol. 5, no. 9/10, pp. 341-345, 2002.

[30] "MARY TTS Home page," 2019. [Online]. Available: https://github.com/marytts/marytts. [Accessed 28 02 2019].

[31] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language,* vol. 19, no. 1, pp. 55-83, 2005.

[32] T. J. Sefara, "GitHub," 2019. [Online]. Available: https://sefaratj.github.io/tts-lid. [Accessed 1 March 2019].