Grammar-driven Text-to-speech Application for Articulation of Mathematical Expressions

Mosibudi Mercy Mogale*, Tshephisho Joseph Sefara†, Tumisho Billson Mokgonyane*
Madimetja Jonas Manamela*, Thipe Isaiah Modipa*
*Telkom Centre of Excellence for Speech Technology, Department of Computer Science
University of Limpopo, South Africa
[1]mogau.mosibudi@gmail.com
[3]mokgonyanetb@gmail.com
[4]jonas.manamela@ul.ac.za
[5]thipe.modipa@ul.ac.za
†Next Generation Enterprises and Institutions, Council for Scientific and Industrial Research, South Africa
[2]tsefara@csir.co.za

***Abstract*—Natural Language Processing (NLP) forms one of the important and fundamental components of speech synthesis while a language grammar forms one of the important requirements for NLP tasks. One of the major requirements in processing speech synthesis tasks is the correctness of grammar analysis. Grammar-based applications tend to be effective when embedded within text-to-speech (TTS) synthesis systems. The TTS synthesis systems assist with the correct word spelling and intonation. Spoken languages plays a vital role to the educational journey of children as their brains are naturally wired to speak but not read and write. This paper presents the development of a grammar-driven TTS application for the reading of mathematical expressions in the Sepedi language. The application front-end component parses mathematical expression text inputs before a TTS synthesis system processes them to produce the correct articulation of the mathematical expression. Acceptable performance results are observed when the application is evaluated using word error rate for intelligibility, and subjective mean opinion score for pronunciation, naturalness, pleasantness, understandability, and overall system impression. The application achieved an accuracy 84,85%.**

***Index Terms*—grammar parser, speech synthesis, language learning, Mathematics**

I. Introduction

Natural languages are built on three different knowledge components: (1) the meaning of words, (2) the sound of words and (3) the grammatical rules according to which words are put together. Language grammar is a formal description of a language often used to recognize language structures like sentences and phrase, syntax and morphology [1]. A grammar specification is an important component of a natural language processing (NLP) application that encompasses checking of phrase and sentence correctness [2]. Grammar-based approaches in online assistive systems have a positive impact on the system performance [3]. There are a number of areas where grammar-based and speech-enabled applications can be used, such as health care services, political events, and education.

The use of technology in education is referred to as computer aided learning (CAL). Usun [4] defines CAL as the communication between a learner and a computer system with an instruction to follow and this involves a computer program or file developed specifically for educational purposes. The use of CAL is rapidly growing in educational institutions ranging from basic to tertiary education teaching and learning [5]–[9]. Educational technologies incorporate tools that enable learners to improve their academic performance. These tools include computer systems, smartphones, and Blackboard learning management systems. In addition, TTS synthesis functionality embedded in most mobile devices and computers can be used to fast-track the learning of additional natural languages. In a multilingual country like South Africa, additional language learning is still a challenge to most learners. Hence, the development of speech-enabled applications as an attempt to augment spoken language processing tools may play a vital role to learners.

As it is an approved policy, the National Department of Education in South Africa offers teaching and learning at foundation phase in home languages and learners have to transition to a national language of instruction at an intermediate phase which is English [10]. This transition leads to poor academic performances for most English second language learners [11], [12]. Lack of support and assistive resources to facilitate efficient mother tongue education by government to enforce policies hamper optimal cognitive development of such learners [13].

Mathematical expressions play an important role in scientific documents by usually being applied in solving problems, using theories in mathematics, physics and other scientific and engineering fields [14]. As a potential intervention strategy to the teaching and learning of mathematics as a subject, this paper presents the development of a grammar-driven TTS application that is capable of articulating mathematical expressions at the foundation and intermediate phase of education in the Sepedi language, which is one of the indigenous official languages of South Africa predominantly spoken in the Limpopo province. Constant development of various text-to-speech systems using one of the indigenous languages of South Africa assist the indigenous speakers to make the best of technology [15].

The rest of the paper is organised as follows. Section II gives an overview of grammar parser and speech synthesis. Section III details the methodology including application architecture, algorithm of the grammar parser, and integration of the grammar parser with speech synthesis. The evaluation procedure is described in Section IV. Section V discusses the results and the paper is concluded in Section VI with the future work.

II. Background

A natural language grammar is an important key component of NLP as it determines the syntactic form of sentences/expressions to be processed. Grammar parser output is very important for day-to-day children education. It is important to make sure that the output produced after all the processes of grammar parsing meets the requirements of grammar analysis. Santaholma [2] presents a methodology of grammar sharing techniques for rule-based multilingual NLP systems such as machine translation systems for recognition, analysis and generation of English, Japanese and Finnish languages.

Embedding a grammar parser in a speech recognition system improves the student's learning performance [16], the same intuition can be utilised by embedding a grammar parser in a speech synthesis system which may enhance the performance of the learners. A TTS engine shown in Figure 1 is made up of two parts: NLP and Digital Signal Processing (DSP) modules. The NLP module contains text, phonetic and prosodic analysis [17].

a) Text analysis: contains three tasks, the first task is the document structure detection module which provides a context for all other modules, the second task is the text normalisation process which converts raw text with symbols like numbers and abbreviations into the equivalent of written words. The third task is linguistic analysis which recovers the syntactic constituency and semantic features of words.

b) Phonetic analysis: assigns phonetic transcription to each word, divides and marks the text into prosodic units, like phrases, clauses, and sentences. Phonetic conversion has two approaches which are rule-based and dictionary-based approaches. A rule-based approach is applied for unknown words while a dictionary-based approach is used for known words.

c) Prosody analysis: is the study of the intonation and rhythmic aspects of language contextual analysis. Prosody can be affected by emotion, mental state and speaker attitude [18]. It determines intonation, amplitude and duration modelling speech.

The Digital Signal Processing (DSP) module contains speech synthesis methods like rule-based (formant and articulatory synthesis) and data-driven methods (concatenation synthesis) to generate synthesized speech [19].

Speech synthesizers are useful when embedded in e-learning tools. Experts have been developing Computer Assisted Language Learning (CALL) for several years now [16]. Human language technologies such as CALL makes things more easier for humans [5]–[7] and speaking practice is said to be important for education especially when learning a second language [21].

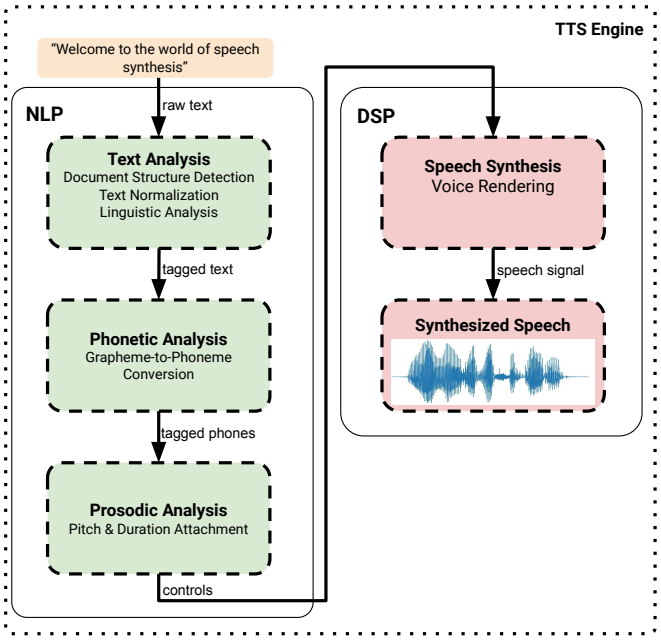


Fig. 1. Basic system architecture of a TTS system [20].

CALL is often a grammar system that focuses on the rules of grammar with respect to the speaking practices [22]. These systems already exists for European languages like Dutch, English, German and Finnish [23] to assist the average and below average learners in learning and mastering the second language. CALL system abilities have shown an improvement in parsing input text [22]. Children from disadvantaged backgrounds have the potential to benefit from computer-assisted instructional technology in the areas of general literacy and mathematics skills [9], [24]. The learners' performance show that assistive computational technology is the best method that can improve learners' understanding of mathematics along with its components that provide active learning [25].

III. Methodology

This section discusses the application architecture, front-end grammar parser and its integration with the back-end speech synthesiser.

A. System Architecture

The overall application architecture is illustrated in Figure 2. Grammar parser has a series of the steps as indicated in the diagram. The first step is lexical analysis where lexer scans every mathematics input and produces a corresponding token of each input. The parser scans the produced token and provides the parser tree which represents concrete structure of words or phrases in a computer and produces parsed results of an input. After the input has passed all the parsing stages then it will proceed to the TTS for the pronunciation.

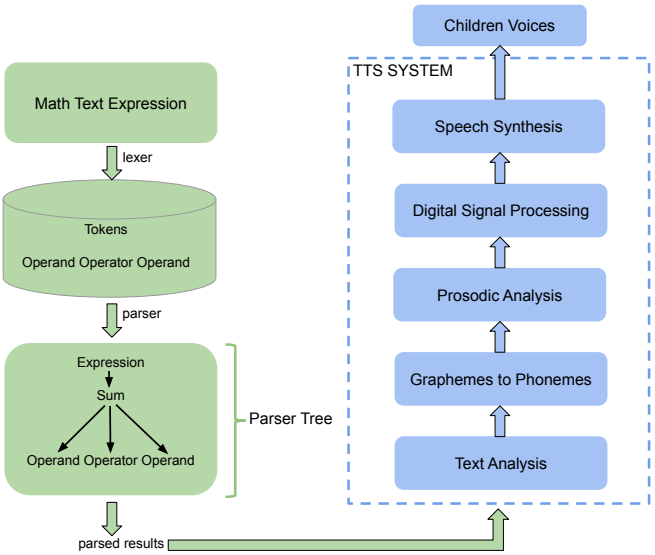


Fig. 2. Flowchart of the proposed application

Algorithm 1: Grammar Parser Algorithm

Input: exp: math expression
Output: $sent$: a sentence
function `GParser(`exp`)`**:**
  Let V be a vocabulary;
  Let E be the splitted exp using white spaces;
  for i **in** E **:**
    if i *is digit* **:**
      $sent+ = digitnormalizer(i)$;
    if i *startswith* R **:**
      $sent+ = currencynormalizer(i)$;
    if i *contains :* **:**
      $sent+ = timenormalizer(i)$;
    if i *endswith* C **:**
      $sent+ = temperaturenormalizer(i)$;
    else:
      $sent+ = arithmeticnormalizer(i)$;
  return $(sent)$;

First step of TTS synthesis is text analysis, a process of normalisation of input text and linguistic analysis where the process analyses the content of the input expression. The second process in TTS synthesis is phonetic analysis which is a grapheme to phoneme conversion where the process uses a set of rules by converting the orthographical symbols into phonological symbols in order to make the pronunciation for each input. The prosodic analysis determines the speaker features. This is a process where we find the pattern of stress, pattern of rhythm and intonation of the text input. DSP is the process which checks for level of accuracy and reliability in speech in case there is a need for improvement. The speech synthesis component forms a set of language units connected to each other continuously to minimise the amount of audible distortion.

B. Front-end: Grammar Parser

A front-end application is composed of a series of scripts containing algorithms that are used to derive the grammar. A $GParser$ function is used to convert symbols and numeric characters into textual representations of Sepedi equivalent form. This function retrieves words from the predefined vocabulary. The vocabulary consists of numbers, symbols, and their word from representation. The algorithm is illustrated in Algorithm 1.

A $GParser$ function depends on other five functions defined as follows:

- $digitnormalizer$ function converts numbers into words.
- $currencynormalizer$ function converts currency into words. The function assumes the first character is a letter (currency symbol) followed by a number. It uses the vocabulary to retrieve the currency and calls $digitnormalizer$ function to convert remaining digits to word form.
- $timenormalizer$ function converts time into words. This function assumes a colon is in the middle. It uses a time template to fill in the numbers which are converted using $digitnormalizer$ function.
- $temperaturenormalizer$ function converts temperature into words. This function is similar to $currencynormalizer$ but in this case a character or symbol is at the end.
- $arithmeticnormalizer$ function uses a vocabulary to convert a given symbol into word form.

C. Back-end: TTS

We adopt the back-end TTS synthesis system that contains South African languages [26]–[28], using Sepedi. We take advantage of the Application Programming Interface (API) available on the back-end TTS to integrate our grammar parser. The integration process using the API is shown in Fig. 3. The grammar parser inputs an equation and outputs a normalised text that is sent to the back-end using HTTP GET API requesting a synthetic voice. The API shown in Fig 3 are explained as follows:

- INPUT_TEXT: represents an input text.
- INPUT_TYPE: represents the data type of the input text.
- AUDIO: represents the type of the output audio.
- OUTPUT_TYPE: represents the data type of the audio.
- LOCALE: represents the language code that is used to produce the synthesised speech.

IV. Evaluation

This section describes Mean Opinion Score (MOS) and Word Error Rate (WER) evaluation methods to test the proposed application.

A. Mean Opinion Score

The system was evaluated by 21 subjects (11 males and 10 females) using the MOS to subjectively measure the quality of the synthetic speech in terms of pronunciation, naturalness, pleasantness, understandability, and overall application

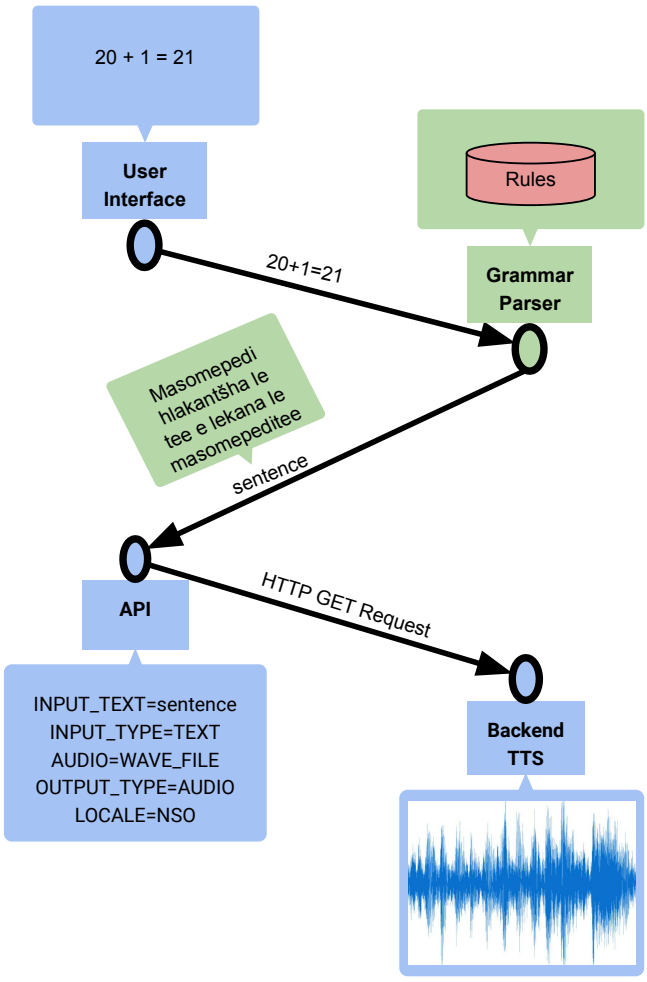


Fig. 3. Integration of the grammar parser with TTS synthesis system

impression. Subjects listened to the six expressions shown in Table I in Sepedi in a sound controlled room. The expressions include arithmetic, currency, year, date, temperature, and time.

TABLE I
GRAMMAR PARSER PARSING THE EQUATIONS

Expresion	After Grammar Parser
$20 + 1$	Masomepedi hlakantšha le tee
$R10 + R5$	Lesome la diranta hlakantšha le ranta tše hlano
$2019 - 1998$	Dikete tše pedi le lesomesenyane ntšha sekete le makgolo a senyane le masomesenyane seswai
$10Days - 3Days$	Matšatši a lesome ntšha matšatši a mararo
$10°C + 10°C$	Kgato tše lesome tša Celsius hlakantšha le kgato tše lesome tša Celsius
$11:00 + 01:00$	Iri ya lesometee hlakantšha le iri ya pele

Subjects were given a questionnaire in a Likert scale of 1 (horrible) to 5 (best) to rate the application. The following questions were asked:

- How would you rate pronunciation of the synthesised speech?
- How would you rate naturalness of the synthesised speech?
- How would you rate the pleasantness of the synthesised speech?
- How much listening effort was needed to understand the synthesised speech?
- How would you rate the overall system impression?

The mean of the responses is calculated to compute the MOS results. The MOS is a performance metric applied to measure the quality of speech and the metric is calculated using the following equation [29].

$$MOS = \frac{1}{n}\Sigma_{i=1}^{n} x_i \tag{1}$$

where x_i is the score assigned by evaluator i and n is the total number of subjects.

B. Word Error Rate

Word Error Rate is a measure of the performance of a system with respect to recognising word sequence that might have a different length from the reference word sequence [30]. The WER 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. Subjects were asked to write down the expressions they heard. We have applied WER on six sentences constructed from Mathematical expressions (see Fig. I). The intelligibility measure is captured by the WER metric calculated as follows:

$$WER = \frac{Sub + Del + Ins}{N} \tag{2}$$

where Sub is substitution, Del is deletion, Ins is insertion, and N is the total number of words. We use this metric to measure the intelligibility of the application.

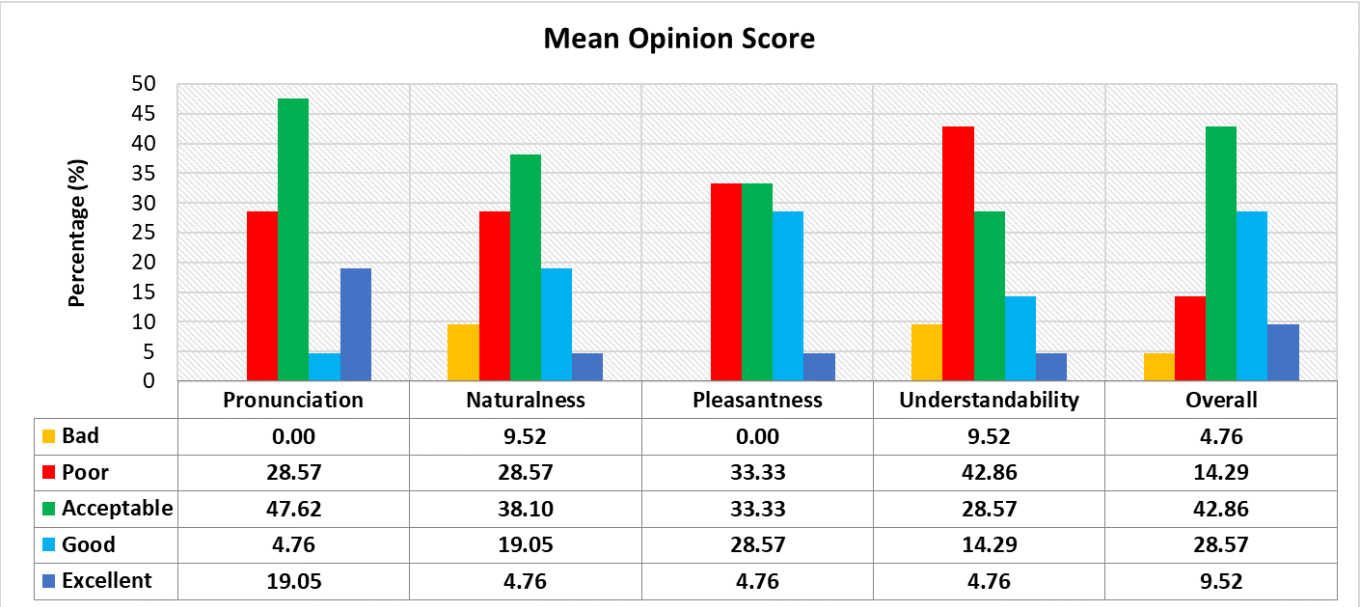
V. PRELIMINARY RESULTS AND DISCUSSIONS

The data collected during the evaluation of the application was analysed using descriptive statistics. Figure 4 shows the ratings for *pronunciation, naturalness, pleasantness, understandability, intelligibility* and *overall application impression*, and the MOS results for the ratings is then shown in Table II.

a) Test for pronunciation: was rated as excellent by 19,05%, good by 4,76%, acceptable by 47,62%, and poor by 28,57% of the respondents. This means that it received an acceptability level of 71,43% and obtained a score of 3.14 which means the application is good at pronunciation.

b) Test for naturalness: was rated as excellent by 4,76%, good by 19,05%, acceptable by 38,10%, and poor by 28,57% and bad by 9,52% of the respondents. This means that it received an acceptability level of 61,91% and obtained a score of 2.81 which means the naturalness of the application is acceptable.

c) Test for pleasantness: was rated as excellent by 4,76%, good by 28,57%, acceptable by 33,33% and poor by 33,33% of the respondents. This means that it received an acceptability level of 66,66% and obtained a score of 3.05 which means the pleasantness of the application is acceptable.



	Pronunciation	Naturalness	Pleasantness	Understandability	Overall
Bad	0.00	9.52	0.00	9.52	4.76
Poor	28.57	28.57	33.33	42.86	14.29
Acceptable	47.62	38.10	33.33	28.57	42.86
Good	4.76	19.05	28.57	14.29	28.57
Excellent	19.05	4.76	4.76	4.76	9.52

Fig. 4. Subjective Evaluation Results

d) Test for understandability: was rated as excellent by 4,76%, good by 14,29%, and acceptable by 28,57%, poor by 42,86% and bad by 9,52% of the respondents. This means that it received an acceptability level of 47,62% and obtained a mean score of 2.62 which means the half of the subjects had difficulty to understand the synthesized speech while the remaining understood the system.

e) Test for overall application impression: was rated as excellent by 9,52%, good by 28,57%, acceptable by 42,86%, and poor by 14,29% and bad by 4,76% of the respondents. This means that it received an acceptability level of 80,95% and a mean score of 3.24.

TABLE II
MEAN OPINION SCORE

Metric	MOS	Meaning
Pronunciation	3.14	Acceptable
Naturalness	2.81	Acceptable
Pleasantness	3.05	Acceptable
Understandability	2.62	Effort required
Overall	3.24	Acceptable

We can observe that the subjects needed little effort to listen and understand the grammar that has been produced from the application and that may have been caused by speaker having different accent than the subjects. Pronunciation tends to differ with accent since accent is a way we pronounce words and it always tends to differ with pronunciation of the speaker. This also affects the naturalness of the speech because the naturalness of the speech is often judged based on the accent and how the speaker pronounces words and sentences. The application was satisfactory to the subjects and gave pleasuring feedback on the system's level of pleasantness. The overall application impression was acceptable and this may imply that the TTS system produces effective and satisfactory results to the intended target group.

We obtained a WER of 15,42% and accuracy of 84,85% for intelligibility and this means that the application may have failed to transcribe some of the words, especially words that are not present in the Sepedi language, such as Celsius.

VI. Conclusion and Future Work

This paper presented a grammar-driven TTS application for articulation of mathematical expressions. We explained the algorithm of the grammar parser for mathematical expressions. We described the integration of the parser with an existing speech synthesis system using API calls. We evaluated the whole application using subjective evaluation method and word error rate. From subjective results, the application is found to have better pronunciation, acceptable naturalness, acceptable pleasantness, listening effort is less, and acceptable overall application impression. From WER results, the system scored an acceptable error of 15.42% which may imply the application is found to be intelligible.

To ensure that the application meets the required expectation the future work will: (i) collect more data recorded by children at the foundation and intermediate phase to create a TTS synthesis system. (ii) add more rules for the parser to be robust on complex expressions.

Acknowledgement

This research project is facilitated at the University of Limpopo Telkom Centre of Excellence for Speech Technology in the Department of Computer Sciences.

References

[1] R. T. Oehrle, E. Bach, and D. Wheeler, *Categorial grammars and natural language structures*. Springer Science & Business Media, 2012, vol. 32.

[2] M. E. Santaholma, "Grammar sharing techniques for rule-based multilingual NLP systems," in *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)*, Estonia, May 2007.
[3] M. Rayner, P. Bouillon, N. Chatzichrisafis, B. A. Hockey, M. Santaholma, M. Starlander, H. Isahara, K. Kanzaki, and Y. Nakao, "A methodology for comparing grammar-based and robust approaches to speech understanding," in *Ninth European Conference on Speech Communication and Technology*, 2005, pp. 1877–1880.
[4] S. Usun, "Applications and problems of computer assisted education in Turkey," *Turkish Online Journal of Educational Technology*, vol. 5, no. 4, pp. 11–16, 2006.
[5] T. Sefara, P. Malatji, and M. Manamela, "Speech synthesis applied to basic mathematics as a language," in *South Africa International Conference on Educational Technologies*, April 2016, pp. 243–253.
[6] P. T. Malatji, T. J. Sefara, and M. J. D. Manamela, "Creating accented text-to-speech english voices to facilitate second language learning," in *South Africa International Conference on Educational Technologies*, April 2016, pp. 234–242.
[7] P. T. Malatji, M. J. Manamela, and T. J. Sefara, "Second language learning through accented synthetic voices," in *South Africa International Conference on Educational Technologies*, April 2017, pp. 106–116.
[8] 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)*, Nov 2016, pp. 47–52.
[9] 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 *2017 IEEE AFRICON*, Sep. 2017, pp. 794–799.
[10] S. Taylor and M. von Fintel, "Estimating the impact of language of instruction in South African primary schools: A fixed effects approach," *Economics of Education Review*, vol. 50, pp. 75–89, 2016.
[11] S. J. Howie, "Language and other background factors affecting secondary pupils' performance in mathematics in South Africa," *African Journal of Research in Mathematics, Science and Technology Education*, vol. 7, no. 1, pp. 1–20, 2003.
[12] A. Mji and M. Makgato, "Factors associated with high school learners' poor performance: a spotlight on mathematics and physical science," *South African Journal of Education*, vol. 26, no. 2, pp. 253–266, 2006.
[13] 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.
[14] J. Zhang, J. Du, and L. Dai, "Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 221–233, Jan 2019.
[15] N. Baloyi, M. Manamela, and N. Gasela, "A text-to-speech synthesis system using hidden Markov models for Xitsonga," in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, September 2012.
[16] M. Rayner, N. Tsourakis, C. Baur, P. Bouillon, and J. Gerlach, "CALL-SLT: A spoken CALL system based on grammar and speech recognition," *Linguistic Issues in Language Technology (LiLT)*, vol. 10, no. 2, 2012.
[17] I. Isewon, O. Oyelade, and O. Oladipupo, "Design and implementation of text to speech conversion for visually impaired people," *International Journal of Applied Information Systems*, vol. 7, no. 2, pp. 26–30, 2012.
[18] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.
[19] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
[20] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.
[21] B. P. de Vries, C. Cucchiarini, S. Bodnar, H. Strik, and R. van Hout, "Spoken grammar practice and feedback in an ASR-based CALL system," *Computer Assisted Language Learning*, vol. 28, no. 6, pp. 550–576, 2015.
[22] E. Rayner, C. Baur, C. Chua, and N. Tsourakis, "Supervised learning of response grammars in a spoken CALL system," in *Workshop on Speech and Language Technology in Education (SLaTE)*, Leipzig, Germany, September 2015, pp. 83–88.
[23] C. Cucchiarini, M. Ganzeboom, J. van Doremalen, and H. Strik, "Becoming literate while learning a second language - practicing reading aloud," in *Workshop on Speech and Language Technology in Education (SLaTE)*, Leipzig, Germany, September 2015, pp. 77–82.
[24] M. H. McManis and L. D. McManis, "Using a touch-based, computer-assisted learning system to promote literacy and math skills for low-income preschoolers." *Journal of Information Technology Education*, vol. 15, 2016.
[25] Z. Yıldız and M. Aktaş, "The effect of computer assisted instruction on achievement and attitude of primary school students," *International Online Journal of Educational Sciences*, vol. 7, no. 1, pp. 97–109, 2015.
[26] T. J. Sefara, M. J. Manamela, and T. I. Modipa, "Web-based automatic pronunciation assistant," in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, September 2017, pp. 112–117.
[27] T. J. Sefara, T. B. Mokgonyane, M. J. Manamela, and T. I. Modipa, "HMM-based speech synthesis system incorporated with language identification for low-resourced languages," in *International Conference on Advances in Big Data, Computing and Data Communication Systems*, August 2019.
[28] T. J. Sefara and M. J. Manamela, "The development of local synthetic voices for an automatic pronunciation assistant," in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, September 2016.
[29] M. Chu and H. Peng, "Objective measure for estimating mean opinion score of synthesized speech," Apr. 4 2006, US Patent 7,024,362.
[30] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson London, 2014, vol. 3.

Mercy Mosibudi Mogale received her B.Sc Honours degree at the University of Limpopo. She is currently studying towards her M.Sc degree at the same institution. Her research interests include speech synthesis, natural language processing, programming languages, and computer-aided learning.

Tumisho Billson Mokgonyane received his B.Sc Honours degree at the University of Limpopo. He is currently studying towards his M.Sc degree at the same institution. His research interests are natural language processing, language learning, machine learning, biometric recognition, programming languages and software development.