Setswana Speech Recognizer for Computer Based Applications

[1] Oratile Leteane [2] Francis, J. Ogwu

Ministry of Education and Skills Development
Gaborone, Botswana

Computer Science, University of Botswana
Gaborone, Botswana

Abstract

This study is the development and adaptation of Setswana speech recognizer into computer applications. Setswana database is used together with sphinx decoder to build a generic Setswana speech recognizer. The recognizer is then adapted into a developed game called General Knowledge Game (GKG) in which the user plays the game using Setswana speech. The developed recognizer's level of accuracy was 60 percent on the worst case. The recognizer improves to more than 80 percent when shorter words are used to drive the application. Participants have shown that they prefer using speech driven applications over traditional approach of using mouse and keyboard. Analysis shows that though it is more effective to use keyboard and mouse to drive computer applications, users still prefer speech interaction because HCI method is easy to learn particularly for the users who are semi-literate and illiterate. It shows that using traditional approach (mouse and keyboard) requires some degree of literacy for someone to be competent while with speech interaction, anyone can use

Keywords: *Setswana, Phoneme, Illiterate, General knowledge Game Database and Phonetic sequence*

1. Introduction

The adoption of spoken language technology as HCI has the basic goal of improving the interaction between users and computers by making computers more usable and receptive to the user's needs. According to [2] a natural language is a desirable component of human to machine interaction. Even though research in using natural language has been successful, it has been biased to English and few other western languages. Indigenous languages especially those spoken in African countries have been neglected although many people speak these languages. Research show that there is large number of population found in remote areas and majority of these people are illiterate [1]. It is important that speech based systems be developed to allow the illiterates to use their mother tongue to operate today's computers. Thus, speech technology needs to be localized so that users can use language they know fluently to drive computer applications. This will reduce the digital illusion problem identified as major problem affecting those who are semi literate and illiterate [1]. Setswana is one of the indigenous languages that is not supported when it comes to HCI. It is a language spoken by around 4.5 million people in Southern Africa. A technology that supports Setswana speaking people to operate computers is needed. Though research carried by [4] was a breakthrough in Setswana speech recognition, it was restricted to numbers and not adaptable. The Setswana recognizer should be adaptable to any speech based computer application. Different researchers may have use different ways of preparing data for the development of speech database. Some consider collecting news papers and getting all words and sentences that appear, which will possible cover larger vocabulary of the language [27]. Some write down different words of the language by asking questions to participants and getting words that they use [28]. For Setswana speech corpus preparation, words relevant to the case study were considered, written down and recorded from different participants. Therefore the objective of the project was to develop an automatic speech recognizer for Setswana that can be adapted to develop computer based applications where interaction is through Setswana speech.

2. Literature Review

Spoken language interface to computers is a topic that has lured and fascinated engineers and speech scientist for decades [2]. Speech recognizing system is a specific form of natural interaction where users speak and listen to an interface rather than type or write on the screen [18]. According to [16], speech recognizing systems have two major components that

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

form its backbone. These components are Automatic Speech Recognition (ASR) and Text To Speech (TTS). Speech syntheses take the parameters from fully tagged phonetic sequence and generate the corresponding speech waveform [24]. ASR is a data and resource intensive process both during training and recognition. Data used for training normally should be from different speakers for better recognition accuracy [7]. Recognized word is an output from speech processing module and may either be a final product or an input to speech recognizing system for further processing. It is observed that the recognition of continuous speech is affected by the rate of spoken language [11]. First, the acoustic realizations of phonemes, the smallest sound units of which words are composed, are highly dependent on the context in which they appear [24]. In [21], SAPI abstracts the developer from the low level details of the SR engine, nonetheless, it is essential that the developer knows the potential, functionalities and work realized by the engine, in order to model and optimize the target application. In the previous researches, most developed and operational voice recognizing applications are based on whole word matching [18]. Modern general-purpose speech recognition systems developed using statistical approaches are based on Hidden Markov Models [11]. HMMs can be trained automatically and are simple and computationally feasible to use [11]. It is solved using Baum-Welch algorithm [15]. Each word, or each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes [25]. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord [20].There are numerous decoders that can be used in the development of speech recognizers with different searching algorithms and the recent once use HMM. Some examples of these decoders include Hidden Markov model Toolkit (HTK) decoder found in [11], Sphinx decoder in [19] and one pass decoder in [22]. The unique features of each system must be studied and carefully weighed against the objectives of the particular task at hand. As discussed in [23], both of these toolkits provide tools that guides in the training of acoustic models. At this point, all possible paths that have reached, the exit state looked at and return the highest scoring path as the result to the application [19]. Existing speech databases [26] and [27] support English language only. Other researchers have developed speech database and plug it in to sphinx-4 framework combining it with other readily made components to come up with speech recognizers of their chosen languages [21]. The speech database should consist of acoustic models, dictionary and language models of interest. Once the database is plugged into the linguist block, the Sphinx-4 decoder is configured towards newly plugged linguist module so that it can use a search Graph module to perform searching process. According to [29], the first step that can be followed in creating a speech database for building an ASR is the generation of an optimal set of textual sentences to be recorded from the native speakers of the language. One major breakthrough in the development of speech database particularly for indigenous languages is through African Speech Technology project [13]. The African Speech Technology project aimed at developing telephone speech databases for five of South Africa's eleven official languages, i.e. South African English, Afrikaans, and three African languages, Zulu, Xhosa, and Southern Sotho. In their research, speech data was collected through phone calls. Currently, there is no existing speech database for Setswana language, thus, in this research the database will be developed from scratch.

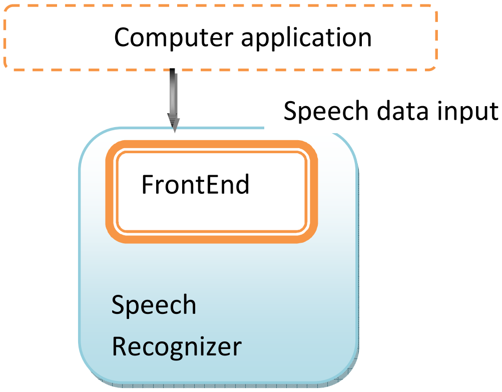
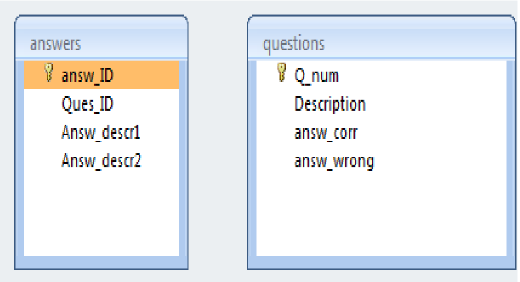


Figure 1: Interfacing computer application and speech recognizer

3. Methodology

The Setswana speech recognizer under development is adaptable to any Setswana based speech driven application for as long as the words needed to drive the application are within defined set of recognizable words in the database. The computer application used as a case study is a computer game called General Knowledge Game (GKF). This is a game consisting of several questions that have two answers. One of the answers is wrong and the other is correct. The user chooses the right answers. This game has two versions; one is developed so that it is played using keyboard and mouse while the other can be played using Setswana speech. The questions and answers are stored in a database and both systems connect to the same database to load questions and answers which are written in Setswana language. Figure 2 show the tables that store questions and answers used in GKG game. The game is timed and when the given time elapses before the answer is provided, the system marks it wrong and the next question is loaded.



answers
answ_ID
Ques_ID
Answ_descr1
Answ_descr2

questions
Q_num
Description
answ_corr
answ_wrong

Figure 2: Tables storing questions and answers for GKG

The game is played using the traditional approach. The user uses a mouse to click and keyboard to type. In this part of the game, at runtime questions and answers are loaded into the application and displayed in graphical user interface. The user selects the right answers by clicking on the right options. If the answer is correct, the score is updated. To build a general purpose Setswana speech recognizer, a huge amount of speech database is needed and it was not feasible in this project due to time constraint. The recognizer was adapted to play a speech driven GKG version of the game as already mentioned. Questions and possible answers were words used to play the game. However, answers coming from the user are actual words that the speech recognizer recognizes during the game

Words:

Seretse, khama, bongwe, bobedi, boraro, bone, botlhano, borataro, bosupa, borobabobedi, borobabongwe, lefela, karabo, ee, ntlha, Gaborone, Botswana, etc

Sentences: Karabo ya ntlha

Motse mogolo wa bostwana e bong Gaborone

Dingwaga di le lesome le borataro

Creating speech data by recording voices of different speakers that was converted into Mel Frequency Cepstral Coefficients (MFCC) was carried out. The creation took into consideration that people may pronounce same words differently. In this research, one word was recorded from at least fifteen different speakers in order to accommodate different pronunciations. The speaker's characteristics were based on their level of literacy, ages, gender, nationality, district they come from and language dialects. All these characteristics were recorded and identified by their name and also given a unique number for their easy differentiation. During recording sessions, each sound wave was recorded with the parameters below;

Sampling rate of the audio: 16 kHz

Bit rate (bits per sample) : 16

Channel : mono (single channel)

In this work, 16 kHz sample rate was chosen because it provides more accurate high frequency information and 16 bit per sample divided the element position into 65536 possible values. Audacity was used for recording and editing of sound files. Recording was done at night; between 6:00 pm and 8:00 pm and early morning between 6:00 am and 7:30am in a quit office environment. This information is provided to the trainer through a file called the transcript file. Transcript file is needed to represent what the speakers are saying in the audio file. Thus, in this file, the dialogues of the speaker are noted exactly the same precise way it was recorded, with silence tag (starting tag <s>, ending tag </s>), followed by the file identity which represent the utterance. Two transcription files were created to represent what was recorded on the audio files. One of them was used to train the recognizer for recorded words and the other mainly for testing purpose. A file called SETS_ASR_train-transcription and SETS_ASR_test. transcription in project folder represents these files. Syntax below represents one line in the transcription file:

A pronunciation dictionary is responsible for determining how a word is pronounced. It has all acoustic events and words in the transcripts mapped onto the acoustic units to be trained. Redundancy in the form of extra words is permitted. To make this file, resources from SETS_ASR pronunciation dictionary is used; if any pronunciation of an entry is not available in the SETS_ASR dictionary then G2P (grapheme to phoneme) module of SETS_ASR is used to generate the sound unit sequence. The file is saved with .dictionary extension. A typical example of pronunciation dictionary is as below:

Motshameko m uh t sha mi k oh

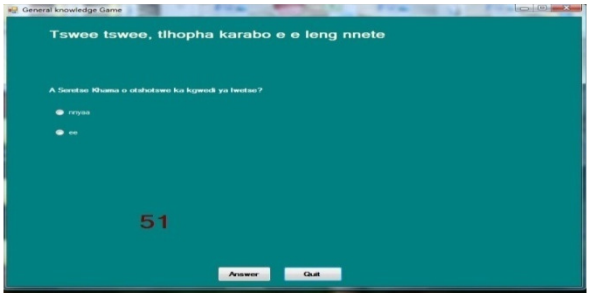
Seretse c ih r ih t hse

3.1 Setting up the System Environment

Perl was used to run the scripts while C compiler was used to compile the source code. The entire training was conducted in Linux environment. CMU Sphinx Train was selected tool for training the speech recognizer under development. Training was carried out in sets_asr folder, which was stored in the same folder as Sphinx Train. The Setswana database was used to train and evaluate speaker-independent Setswana speech recognition systems for various acoustical conditions. To nurture the development of speaker-independent systems, the Setswana database was collected using a large number of speakers selected from staff Ministry of Education staff to represent Botswana population. Setswana speech database was adapted into a Setswana game that is played using Setswana voice.

3.2 Speech vs. Mouse-Keyboard GKG

The game called General Knowledge Game (GKG) and was developed in java programming language. GKG consisted of seven different general knowledge questions. A user is given ten seconds to answer a question. If ten seconds elapse before the user answers the question, then a user is marked wrong. User's input is through either traditional interaction, which is mouse and keyboard or Setswana speech. To play GKG using keyboard and mouse, the question is written on screen and all possible answers are displayed as well. User chose the best answer based on their knowledge and personal experience. Figure 3 shows GUI interface that the user interacts with when playing the game using mouse and keyboard. On the top part is the general question that the user has to answer. In the middle are alternative answers that the user has to choose from. Then on the bottom is timer counting down. When the timer reaches zero before the answer is provided, then the player is marked wrong and next question is loaded. On the other hand, the user played the game using voice. Setswana words were uttered into the microphone. The developed Setswana recognizer picks the uttered words, converts to text and sent it into the GKG game. If the word is the correct answer, the score is update respectively. Figure 4 shows the screen that the user interacted with when playing the GKG using Setswana voice. For the two methods, all variable are the same except that the other is driven by voice and the other driven by mouse and keyboard.



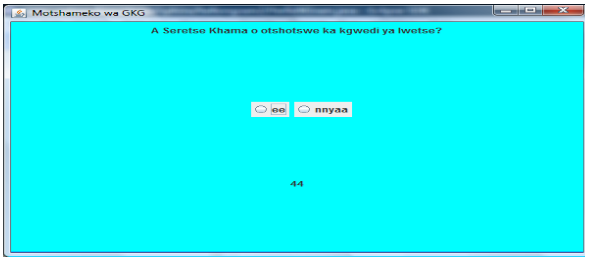
Tswee tswee, tlhopha karabo e e leng nnete

A Seretse Khama o otshotswe ka kgwedi ya lwetse?

● nnyaa

● ee

Figure 3: GUI for playing GKG using mouse-keyboard interaction



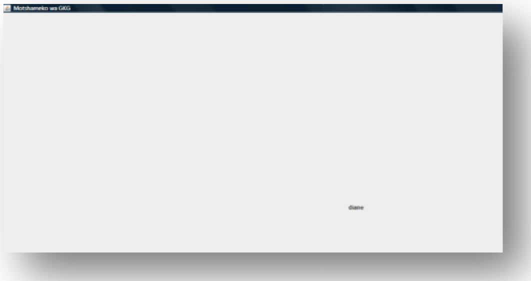
A Seretse Khama o otshotswe ka kgwedi ya lwetse?

○ ee ○ nnyaa

Figure 4: GUI for playing GKG using Setswana speech

3.3 Participants for recognizer accuracy

In the first experiment, participants were chosen from street around Gaborone. The experiment ran for five days.. The speech interface for testing the accuracy of the system is shown below. The interface was developed using java programming and accepts words in speech and output the words in textual form. Figure 5 shows a word "diane" displayed after participant uttered the word. All participants were trained on how to use the system. The users were advised to speak at average speed because most of research emphasized that speech systems perform better when words are spoken at average speed [5, 10].



diane

Figure 5: Text "diane" displayed after participant uttered the word dia

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

A population of fifty people from different industries in Gaborone was used to represent population of users. The group was divided into two groups of twenty five. The first group started playing the game by interacting with computer using traditional approach of mouse and keyboard and then using speech after. The second group is by means of speech and then mouse and keyboard later.

To record the accuracy level of speech recognizer, observation was considered and chosen as an approximate method. Users were given list of Setswana words to read to the system. On recognition, the words were output to the computer screen. Each word was said several times by different speakers and number of times a correct word output was recorded. The data gathering technique used was questionnaire. Participants were given a likert scale with questions rating from 1 representing disagreeing to 5 representing strongly agreeing

4. Analysis

The recognition performance evaluation of an ASR system was measured on a corpus of data different from the training corpus. A separate test corpus, with new Setswana language records, was created as it was previously done with the training corpus. The test corpus was made of 50 recorded and labeled data which were later converted into MFCC. In order to test for speaker independency of the system, some of the participants who did not participate when recording corpus for training were used in creation of the training corpus. To further test the system on live data and also again test its speaker independency, the system was tested by running it live using sphinx-4 decoder.

The recognizer was tested using 20 different speakers who did not participate when recording speech corpus for training and testing. Here the target was to test the accuracy level of the system. Users were given a list of Setswana words to pronounce. Each word was pronounced 20 times. The number of times the system recognized properly or wrongly was recorded. The average accuracy was calculated and was found that the system recorded 60 % accuracy. From the results obtained, it shows that some words recognition accuracy is better than others. The Setswana word “a” has very high accuracy level such that when pronounced 20 times, the possibility of the system picking it is 70%. It was also found that longer words such as “legae”, and “mmm” recorded low recognition accuracy below 50%. In average, it can be concluded that shorter words are recognized better than longer words. It was found that accuracy improves with experience of using the system.

Characteristics of participants are also recorded to ensure that during testing there is balance of age, gender, region of origin, education level and computer experience. Users started the GKG game and uttered the answers to the system. The system responded back by writing on the screen the message: mmm! O botlhale, karabo ya gago e nnete! (congradulations! your answers is correct) or Ka maswabi, o fositse (Sorry, your answer is wrong).

5. Results

These results show that the system is speaker independent. The results also show that it is more effective and efficient to use traditional method of computer interaction to play GKG game over using Setswana voice. The mean for using Setswana voice is 2.5 with standard deviation of 1.27. While on the other hand, the mean of using mouse and keyboard is 3.9 with standard deviation of 1.22. The efficiency of the voice operated system is 2.78 with standard deviation of 1.39 while the efficiency of mouse and keyboard operated system is 2.41 with standard deviation of 1.26.

Participants rated voice interaction method to be easier to learn as oppose to that of using keyboard. The voice based method scored the mean of 2.75 with standard deviation of 1.41 while mouse and keyboard based method scored mean of 2.55 with standard deviation of 1.19. It also shows that it is easier to remember how to use the voice operated system than using mouse and keyboard. Remembering how to use the system scored a mean of 3.03 with standard deviation of 1.44 for voice based interaction method. On the other hand mouse and keyboard method has mean of 2.78 with standard deviation of 1.16.

The participants rated voice based method to be more satisfying than using traditional method of keyboard and mouse. Voice method got mean of 3.05 with standard deviation of 1.36 as opposed to keyboard and mouse which has 2.5 and standard deviation of 1.15. Most of the participants showed more interest in interacting with computer using their mother tongue Setswana. It is possible from these results that people with low level of literacy will develop interest in using computer applications when interaction is through speech and their mother tong. The speech based system is more enjoyable with the mean of 3.3 and standard deviation of 1.45. Participants were highly motivated playing the GKG using speech.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

Mouse and keyboard game has a mean of 2.15 and standard deviation of 1.2

Participants felt that they were more engaged when playing the game using speech over mouse and keyboard. The mean for speech based approach is 3.4 with standard deviation 1.57. In contrary, Keyboard and mouse method scored mean of 2.65 and standard deviation of 1.35.Both approaches were rated to be fun with all means above 3. Voice based approach has mean of 3.25 with standard deviation of 1.48 while mouse keyboard has mean of 3.05 and standard deviation of 1.67.

6. Conclusion

The application was developed such that in one version, users interact with GUI using computer mouse and keyboard and the other is through Setswana voice. This represents a practical exercise that precedes the localization of applications to allow illiterate users to be able to interact with the system.

Experimental testing was carried out in a quite small quiet office environment and same microphones were used. The recognition performance degraded dramatically as we went into real-time testing but was higher when using recorded data. A poor-quality microphone can have large effect on the rate of recognition. Mainly, two types of noise are experienced in such a real-time application. Bad microphone can introduce additive and convolution noise. There are many sources that add to this distortion such as room reverberation which is represented in the speech utterances bounced off the walls and added at the microphone to the original speech signal. Moreover, the speech signal can be significantly distorted depending on the type of microphone transducer used and its mounting These experiments showed that spoken language interfaces can provide easier and more efficient interaction with computer application.

References

1. Katongo, A. Morakanyane , R. 2009, ‘Representing Information for Semi-Literate Users: Digital Inclusion Using Mobile Phone Technology’, Prato Community CIRN 2009 Conference, Botswana

2. Long, B 1994, Natural Language as an Interface Style, Dynamic Graphics Project, University of Toronto, Viewed 20 September 2011. <http://www.dgp.toronto.edu/people/byron/papers/nli.html>

3. Resch, B. 2003, Automatic Speech Recognition with HTK. A Tutorial for the Course Computational Intelligence, Signal Processing and Speech Communication Laboratory, viewed 07 January 2011. < http://www3.spsc.tugraz.at/courses/scl/download/ASR.pdf>

4. Mamadisa, K. 2005. ‘Numeric string recognition in Setswana’. Southern Africa Telecommunication Networks and Applications Conference (SATNAC), RSA

5. Muhirwe, J 2005, ‘Automatic Speech recognition: Human Computer Interaction for Kinyarwanda language’, MSc Thesis, Makerere University, Rwanda.

6. Liu, F 1994, ‘Environmental Adaptation for Robust Speech Recognition’, PhD Thesis, Carnegie Mellon University, Pennsylvania.

7. Wikipedia, 2011, Speech Recognition, Viewed 20 December 2010. <http://en.wikipedia.org/wiki/Speech_recognition>

8. Bugmann, G 2003, Challenges in verbal instruction of domestic robots. Robot intelligence laboratory, University of Plymouth

9. P. Firtpatrick, P and Breazeal, C 2001, Characterising and processing Robot-directed speech, Cambridge: AI lab

10. Angilingua, 2011, Angilingua Speech Recognition, Viewed 15 February 2011

    <http://www.agilingua.com/en/products_voice/overview.php>

11. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P 2002, ‘HMM Definition Files’, The HTK Book, Version 3.1, Cambridge , PP 91-110.

12. Jensson, A 2006, Study on Development of Speech recognition System Using a sparse training Corpus for an Icelandic Dialogue system, M.Sc. Thesis, Tokyo institute of Technology, Japan.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

13. Roux, J 2005, Results of African Speech technology, Viwed on 5 June 2011 <http://www.ast.sun.ac.za>

14. Ogwu, F, Talib, M and Odejobi, O. Text-to-speech processing using African language as case study. Department of Computer science, UB (Botswana) and Aston University (United Kingdom). Taru publishers.

15. Juan, B and Rubiner, L 2007, 'Hidden Markov for Speech recognition', Technometrics,Vol. 33, No. 3, pp. 251-272.
16. Carnegie Mellon University, 2000, Sphinx Train Documentation, Viewed 13 March 2011. <http://www.speech.cs.cmu.edu/sphinxman/scriptman1.html>

17. Ibrahim, A 2009, 'Distributed Speech Recognition over IP Network using Java', MSc Thesis, University of East Anglia, Anglia
18. Preeze, J, Sarp, H and Rogers, Y 2007, 'What is Interaction design', Interaction design- beyond Human computer Interaction, 2nd edition, John Wileys & Sons Ltd, England, pp 2-41.
19. Lamere, P, Kwok, P, Walker, W, Gouvea, E, Singh, R, Raj, B, Wolf, P 2003, 'Design of the CMU Sphinx-4 Decoder', 8th European Conference on Speech Communication and technology (EUROSPECH).
20. Doe, H 1998, 'Evaluating the Effects of Automatic Speech Recognition Word Accuracy' M.Sc. Thesis, Virginia Polytechnic Institute and state university, Virginia.
21. Chowdhury, S 2010, 'Implementation of Speech recognition for Bangla', B.Sc. Thesis, Brac university, Bangladesh
22. Pylkkonen, J, 'An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition', In proceedings of 2nd Baltic conference on human language technologies, 2005, pp 167-172.
23. Novak, A, Dixon, P, Furui, S 2010, 'An Empirical Comparison of Sphinx and HTK models for Speech Recognition', The 18th Indonesian Scientific Conference, 2010,Japan
24. Freitas, J 2007, 'Spoken Language Interfaces for Mobile Devices', M.Sc. Thesis, Instituto Superior de Engenharia de Lisbon, Portugal.
25. Saha, G, Yadhanandan, S 'Modified Mel-Frequency Cepstral Coeffient', International Association for Science and Technology for Development (IASTED), 2045, Austria.
26. Carnegie Mellon University, 2000, AN4 Speech Database, Viewed 13 December 2011. < http://cmusphinx.sourceforge.net/wiki/sphinx4:an4>

27. Paul, D, Barker, J 1992, 'The design for Wall Street Journal-Based CSR Corpus', Proceedings of the workshop on Speech and Natural Language, 1992.

28. Hamdani, G, Selouani, S, Boudraa, M, Algerian Arabic Speech Database (ALGSD): Corpus Design and Automatic Speech recognition Application

29. Anumanchipalli, G, Chitturi, R, Joshi, S, Kumar, R, Singh, S, Sitaram, R, Kishore, P. Development of Indian Speech databases for Large Vocabulary Speech Recognition System

Biography

Mr Oratile Leteane obtained B.Sc. degree majoring in Computer Science and Applied Mathematics at Nelson Mandela Metropolitan University (NMMU) in South Africa in 2007. Then, he obtained M.Sc. degree in Computer Science at University of Botswana (2012). Moreover, Leteane has worked as an assistant lecturer at ABM University College where he taught programming and systems analysis modules for one and half years. To date, Leteane is working as a systems analyst at Ministry of Education and skills development where his key responsibility is to support the development of business support systems for the ministry. Interests include research in development of advanced speech recognizing systems

Francis Joseph Ogwu has a B.Sc. Degree in Computer Science in 1981, M.Sc. Degree in Computer science Communications in 1984, a PhD in artificial intelligence Computer science in 1990 from Obafemi Awolowo University Nigeria. He worked as a lecturer in the above University for 18 years and he is presently working at the University of Botswana as a Professor of Computer Science. He has published with many journals in Europe, America, Asia and Africa. He has attended and chaired most conference round the world. He is a Special member of IASTED (International Association of Science and Technology for Development)