SINHALA SPEECH RECOGNITION SYSTEM FOR JOURNALISTS IN SRILANKA

SINHALA SPEECH RECOGNITION SYSTEM FOR JOURNALISTS IN SRILANKA

M.I.M. Aathif
Department of Information Technology
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
IT16088788@my.sliit.lk

W.M.P.N. Warnasooriya
Department of Information Technology
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
IT16133082@my.sliit.lk

A.M.S.R. Bandara
Department of Information Technology
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
IT16121102@my.sliit.lk

I.M.C.U.Herath
Department of Information Technology
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
IT16113732@my.sliit.lk

Abstract-- Speech and sound recognition systems are more significant in the market today that many of them hold the ability to identify interactive speech. Such devices or software detect and comprehend a voice signal to a text. Speech recognition traces the critical themes as in navigation, fetching, and higher-level inference on audio data. Accordingly, numerous significant advances have been reported in the field of inaccuracy and robustness in journals and conferences. This research focuses on the aspects of improving the audio quality using several noise filtering techniques (Noise cancellation), human vocal separation, speaker diarization, converting the audio to a text, and translating the audio to a Unicode text and grammar correction and also summarization according to different aspects. One of the primary aims of this research is to find reducing unwanted sound effect using machine learning, minimize the time gap of converting an audio to a text that will enhance the efficiency for professionals in fields like journalism in Sri Lankan context. Peoples working in music industry who are looking for efficiency audio data editing tools (vocal separation, Background music removing). This also contributes to a significant social aspect of the people who are suffering from Dyslexia, which is a learning disorder and to aid them in overcoming the interruptions potentially.

Keywords— Speech recognition, Speech-to-text, HMM, HTK, GMM, Speaker Diarization, Application Program Interface, Artificial intelligence, FFT, Python Library, Hidden Markov Toolkit

I. INTRODUCTION

One of the first circulations of machine-controlled speech recognition system is to transcribe speech documents like talks, interviews, lectures, and broadcast news. This paper describes a remote control computer system using speech recognition technology in mobile devices for the blind and physically disabled population. Many people feel difficult and frustrated using machines using a keyboard and mouse. In addition to this probes how the speech recognition system is implemented for a Sinhala Audio to a Text which will be catering to the significant

Importance with the people with hearing and learning disabilities that shows the difficulty in reading due to problems determining speech sounds and learning how they relate to letters and words (decoding). The literature associated with the research manifests a few implementations associated with this study are accessible. with generating positive results in the field. The methodology is woven around four essential parts, and the rest of the discussion declares the experiments and results.

II. LITERATURE REVIEW

A. Identifying the sound environment

Under this section, many scholars have stated the key themes, and they are brought out in terms of de-noising. The removal of A.W.G.N. "additive white Gaussian noise" is primarily done. The authors concentrated on distorted, white-noise audio signals [1]. White noise is complicated to eliminate, as it is present at all frequencies. The authors used a Discrete Wavelet Transform (DWT) to convert noisy wavelet-domain audio signals. Signal was thought to be expressed by DWT coefficients of high amplitude, and ratios of low amplitude represent noise. The thresholding of factors is used to get audio signals with less noise, and they are converted back to the time domain. The authors proposed to adjust universal coefficient thresholding, resulting in a better audio signal. [2]. Li Zhang, Xiaomei Chen, and Bo Zhong – have been learning a great deal about the impact of hearing loss and hearing aid. Leading technology affecting hearing effect is noise control technology. Noise mitigation output adversely impact speech intelligibility, including the physical and mental health of people who have an impaired or defective hearing. A first acoustic signal will be obtained in this process through the experiment device that can replicate actual working conditions. The signal-to-noise ratio (SNR) and the segmental signal-to-noise ratio (segSNR) of the signal will be determined after aligning the output signal and input signal to test the efficiency of hearing aids in noise reduction [3].

B. Speaker diarization

Speaker Diarization's first ML-based works started around 2006, but significant advances only started around

2012 (Xavier, 2012), and it was considered an exceedingly tricky task at the time. Many of the approaches back then were based on G.M.M.s or HMMs (such as J.F.A.) that had no neural networks involved. A tremendous breakthrough came with the release of L.I.U.M., an open-source program that was written in Java dedicated to speaker diarization. There was, for the first time, a widely available algorithm that could perform the function with excellent precision. L.I.U.M.'s main algorithm is a complex process integrating G.M.M. with I-Vectors, a system that used to provide state-of-the-art results in voice recognition tasks. There is usually no background knowledge about the number of speakers involved or their identities. Estimating the number of speakers is one of the key obstacles to the process of speech diarization [4]. To sum up, this function consists of identifying the turns of the speakers, grouping the homogeneous segments of the speakers into clusters, estimating the number of speakers involved in the Text. Classical methods for speaker diarization deal with these three points successively: first finding the speaker turns using the symmetric Kullback Leibler, the Generalized Likelihood Ratio, or the distance methods of the Bayesian Information Criterion (B.I.C.), then grouping the segments during a hierarchical clustering process, and finally estimating the number of speakers afterwards. Before exploring various techniques for the combination of the systems, both diarization approaches and acoustic segmentation were developed independently. The settings of each of them differed as acoustic features or methods of learning and came from experiments conducted over a similar corpus of growth [5,6].

C. Conversion of Audio (speech) to Text

Real-time translation of discrete expression from Sinhala to Unicode text. This study paper, written at the University of Sri Jayawardenepura by M.H.K. Gunasekara and R.G.N. Meegama. They developed a program for translating Sinhala Speech into Unicode text. As their speech recognizer, they used the Hidden Markov Model (HMM) and related Hidden Markov Toolkit (H.T.K.) techniques [7]. Julius decoder, a topology of three states Baking H.T.K., is used by them for the construction of the acoustic model. In a single speaker testing session, in both speakers based and independent stages, they got 85 percent of average accuracy. Their performance assessment shows this program capable of doing the conversion process for both quiet and noisy environments. A significant technique for translation of Bengali Language in real-time Speech to Text is cautious here. This research paper discusses the Bengali language in real-time Speech to Text Translation. It will help us to convert Speech in the Sinhala language to Text. They use the open-source Sphinx 4 platform implemented by Java and provide procedural coding tools to build a custom language acoustic model. Speech to Text conversion is a method of dissecting captured audio phonemes and Translating it to text. Authors developed a language model of their own and a dictionary to translate spoken Bengali words into modern "Bengalish" use We substitute "Bengalish" back to Bengali, after creating the language model and dictionary. The device converts Bengali to Unicode text after it. Kukarella-Text Converter Audio. This software application converts the Audio to Text online [8]. Kukarella provides users with three key features, such as user can use system microphone and convert audio to live text, user can upload the recorded file as well, and the user can enter YouTube and Vimeo URLs to save. Kukarella can convert audio to text in Sinhala language. This conversion was made using Google API. Yet it's not ideal for 100 percent jobs. This software offers the captured audio files a limited period (without login, they provide the sound period of 3min) [9].

D. Grammar Correction

The University of Moratuwa 's project "ෙෙෙෙෙෙේ' ha introduced necessary grammar scenario testing for just three words sentence. Because of the broad reach of the Sinhala language, it is challenging to make more changes to that structure because it is a ruled-based system. Spell and Grammar Testing Tool for Sinhalese; A collaborative work by UCSC, UOM, and SLIIT scholars [10]. Here they mainly concentrated on checking spell rather than checking grammar. Author used techniques such as the interpretation of natural language, machine learning. As the program focused primarily to spell checking, the functionality of spell checking was introduced in an already existing framework, a data-driven method in "ෙෙෙේ' (SUBASA) [11].

III. METHODOLOGY

A. Participants

Researchers were assiting as participants for the module with other team leaders. A population of 4 Males between 22 and 25 years old. The participants used a Python library to convert the Sinhala Unicode text. The objectives are distributed between the four researchers in developing a machine learning algorithm identifying the sound effects inside a recording. The sound environment and the required sound frequencies are further analyzed in terms of extracting the specific sound wave. The converted text is analyzed by the participants to generate a perfect output for the library. The samples are tested on various backgrounds to get a precise result with high accuracy.

B. Apparatus

Google Cloud Speech API: Through implementing powerful neural network models in an easy-to-use Interface, Google Cloud Speech API gives developers to transform audio to texts. To assist the global user based the API identifies over 80 languages and variants. You can transcribe the users' text dictating to the microphone of an application, allow voice command-and-control, or transcribe audio files, among many other use cases. Recognize submitted audio in the file, and integrate with your Google Cloud Storage audio storage, using the same technology that Google uses to drive its own goods.

Natural Language toolkit: NLTK, is a suite of open source software modules, tutorials and issue sets that incorporate ready-to-use courseware for computational linguistics. NLTK covers natural language symbolic and mathematical analysis, and is interfaced with annotated corpora. Students increase and replace existing components, learn structured programming through example, and from the outset they manipulate sophisticated models to process English written in the Python programming language.

Librosa: Python software for manipulating the audio and music signals. At a high level, Librosa provides implementations of a number of common functions that are used in the music information retrieval area. A brief description of the functionality of the library is given in this document, along with descriptions of the design goals, software development standards and notational conventions. In general, the functions of Librosa tend to expose the caller to all applicable parameters. Although this allows experienced users tremendous versatility, it can be daunting for inexperienced users who simply need a clear interface to process audio files.

IV. DESIGN AND DEVELOPMENT

Automatic voice recognition (ASR) and machine translation (MT) pipeline of popular speech-to-text translation systems are some of the few methods. But high-quality ASR needs hundreds of hours of transcribed audio. In contrast, high-quality MT needs millions of parallel text words — resources usable for just a tiny fraction of the approximate 7,000 languages worldwide. Only a few languages have developed speech recognition systems, and Sinhala speech recognition systems are tough to come across. Recognition or recording of a person's identity is the question of 'Who talked when.' The Speech Recognition System allows people, including journalists, writers, students, etc. to accomplish their life tasks and support them in work and educational activities and guides them to perform their functions and make life more comfortable adequately [12].

Most speaker recognition systems are equipped to recognize English only, making it impossible for people to participate in activities that only include the Sinhala language. Often, diarization with speakers may benefit adults/students who have Dyslexia. The whole designing process bases on the system overview and system architecture, which includes all the sub breakdowns related to the central research theme.

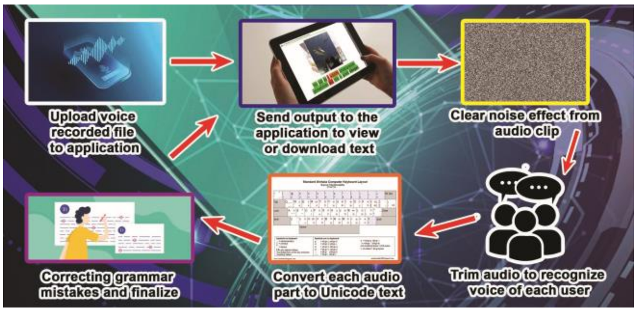
Upload voice recorded file to application
Send output to the application to view or download text
Clear noise effect from audio clip
Trim audio to recognize voice of each user
Convert each audio part to Unicode text
Correcting grammar mistakes and finalize

Figure 1: System Overview diagram

A. Removing the background noise and amplifying the Audio

It is tough to record pure audio with many background noises. The noise and audio will get interrupted together, mixed, and the original sound wave will decrease the clarity. Every intervening auditory signal, such as background noise, acoustically masks and interferes with the intended audible signal in a listening environment. However, noise does interrupt not only peripheral processing but also central auditory processing. CAP includes neural processes necessary for the classification, identification, ordering, grouping, and localization of sensory signals. The identification of the sound effects inside the audio file is the previous step. It is analyzed and divided using FFT Algorithm [13]. Then the sound which can occur in a specific setting can be described. We are using unsupervised learning methodology in here [14]. Its frequency can define every effect of sound. Complete data on rates will be focused. Therefore, the best matching frequency can categorize a specific environment. The noise filtering is done next, which is cut off audio limited to a particular level of rate. The process is thede-noising.

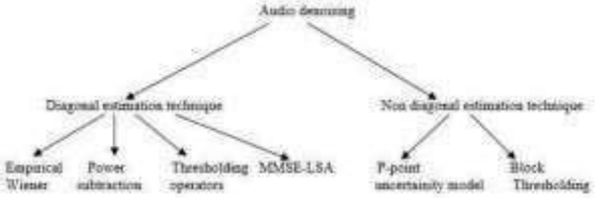

Figure 2: Audio denoising taxonomy

In the noise filtering, a selected frequency range is marked, and it is added with a bandpass filtering technique to trim the audio. In this research, another methodology is focused on the combination of the time domain and time-frequency domain. The method trains a deep neural network to fit this signal, due to a noisy audio clip. Because the fitting is only partly effective and retains

The underlying clean message better than the noise, the network 's output helps to disentangle the clean audio from the rest of the signal. The method is unattended and trains only the specific audio clip denounced [15]. For the final result generation, the stage of cost reduction operation time is considered. The mechanisms like converting stereo to mono, lessening the audio sample rate, and lowering the audio bit depth are verified in this stage. Both methodologies are put into action to sharpen the final output.

B. Speaker Diarization

The scientific community has developed research on speaker recognition/diarization in several different domains over the last few years, with the aim of being usually dictated by funded research projects. From early working with telephony data, broadcast news (BN) became the main focus of research with regards to the late 1990s and early 2000s, and the use of speaker recognition was planned to automatically annotate television and radio transmissions that are broadcast daily around the world. Annotations included automated speech transcription and meta-data marking, including the diarization of speakers [16]. The anatomy and the infrastructure of diarization are intensified in terms of Speech Detection using Voice Activity Detector (VAD) for the noise cancellation and non-speech areas. In Speech Segmentation, extraction is done to pick the short segments from the main sound file ad it to run the LSTM network to generate D vectors for a sliding window.

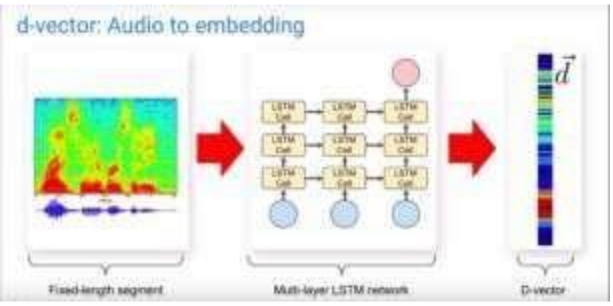

Figure 3: Generating D-vector from segments

The recognition of the audio is done by enabling the embedding extraction for each segment of the sound file, and it aggregates the sections that are belonging to the same data. It is the process of reduction of dimensionality by which an initial set of raw data is reduced to more attainable groups for processing. Extraction of features is the term for methods that select and combine variables into elements, effectively decreasing the amount of data to be processed while still wholly and accurately describing the original collection of data [17]. The clustering of speakers is the job of separating speakers in a database. Such a way, such audio recordings, the aim is to address "who spoke when." Using well-known techniques such as Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) [18], feature extraction directly from the recording is a standard method used in the industry. It determines the number of speakers along with their time stamps.

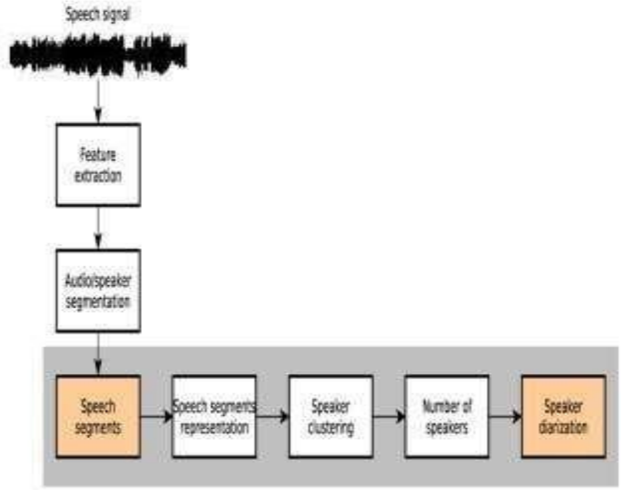

Figure 3.1: Speaker recognition/diarization process

C. Conversion of Audio signal to the Text

Parameterization of speech includes of transforming the voice signal into a set of feature vectors. The purpose of this transformation is to obtain a new representation that is more compact, less redundant, and more suitable for statistical modelling and distance or other score calculation. Many parameterizations of speech used in speaker testing systems are based on a cepstral representation of speech [19]. Here the audio input file is converted to a text format. The problems will arise while the process is ongoing. One of the problems is that all redundant words should be included in the translated text that it should not be unsuitable, inappropriate, and the Sinhala Unicode texts will be displayed in the production. The explanation for this is the device will have the correct word for the audio sound, but it is not the actual output (text) the author expect, and the problematic area should be solved. The thing is to improve our algorithm and maintain a database to store these words and train those words to obtain high precision.

Another question is the English word suggestions. Sometimes the uploaded audio file contains few English words, and our system never knows the corresponding Sinhala Unicode text for it so that system cannot convert those words into text. It then prints the word as production in English. Author also need to handle these kinds of mistakes. So we have to keep a dictionary for handling this case. Then our program can easily translate words from English to Sinhala too. Punctuation marks (./,!?) "are another case. The bulk of web applications that deal without the punctuation marks. Because of this purpose, the user is unable to recognize the sentences, separations, and other issues. The main focus of this research is focused on this stage that carries the essence of the research article. Here the speaker recognized audio signal is converted to the Sinhala Unicode text. The tool kit that is used here is the Python library and done in several stages following the languages' syntactical structure, the phoneme representation. The audio sound filter Kalman is used to verifying the filtration if necessary. The spoken words are then identified using the HMM (Hidden Markov Model), and then the algorithm is used to process the convert to the Sinhala Unicode text [20]. Sinhala words are detected and converted to the Sinhala texts and to represent Singlish words, and a phoneme table is further created and convert that into a Sinhala Unicode text. According to the Sinhala computational grammar [21] laws, the subject, verb, and object of the sentence will be defined along with the sentence form. The singular or plural phrase will then be identified in advanced grammar checking the active or passive voice of the sentence, masculinity or femininity of the sentence. The entire paragraph will be listed as mentioned above and will access every paragraph sentence. The system will automatically identify the existing / non-existing grammar rules and words by using a time-to-time update dictionary. The rest API was introduced during the last stage of deployment. Having the program more available to the users and having it conveniently compatible with the group members' other research components.

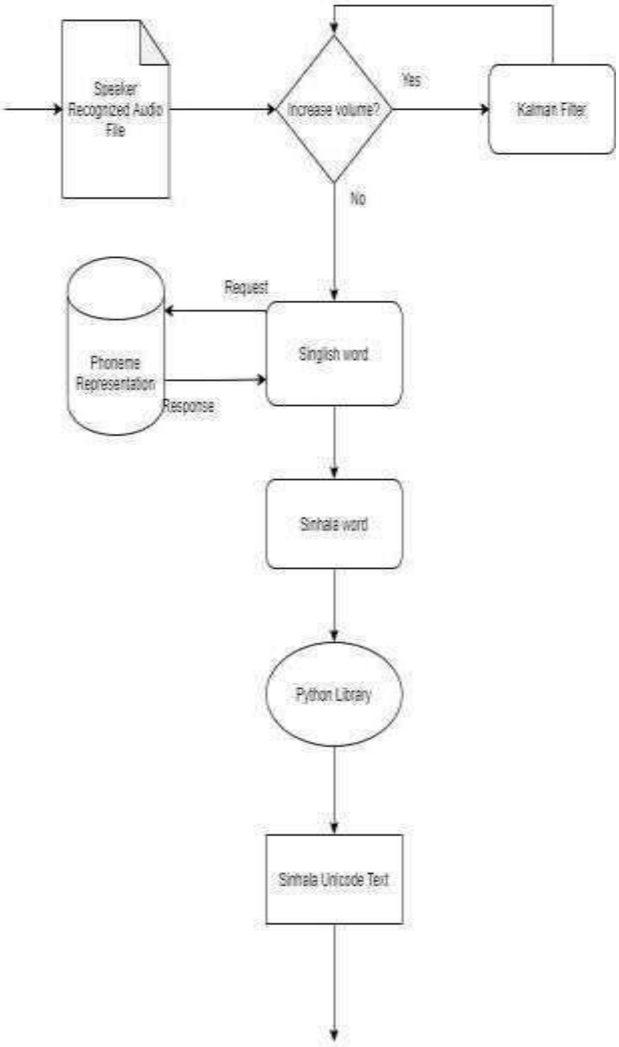

Figure 4: Conversion Process

D. Grammar Correction

The project outcome will be useful for increasing the efficiency of clerical activities at the locations mentioned above. In addition, we have described our target audience as newspapers, government sectors, school kids, teachers, and all individuals who wish to speak Sinhala. As Sinhala, the language in Sri Lanka is the constitutionally recognized official language. The expected outcome can be used by end-users in the real environment. The long-term outcome effects would lead to the preservation of our mother tongue for the good of the generation to come. The field of study that focuses on human language-computer interactions is called Natural Language Processing or NLP [22] in short.

It lies at the intersection of the linguistics of computer science, artificial intelligence, and computation. One of the simple Natural Language processing software or tools for testing language syntax is NLP way for code interpretation, comprehension, and A grammar checker. The field of natural language processing in an Asian language is relatively new, and a lot of tools are yet to be developed. A Grammar Checker is one of these. The rule-based and machine-learning techniques were commonly used in the development of grammar checkers. Because the field of natural language processing for Asian languages is limited, along with the complex grammar, in developing a grammar checking system for Sinhala languages present some problems. These systems were developed using different types techniques and methods.

The main idea of this section is to test the consistency of the grammar of a word phrase or the article. Collecting a suitable collection of data is a major problem for a low-resourced language such as Sinhala. To sort out this problem, Author decided to use the Machine Learning Paradigm Active Learning methodology that is applied to many tasks, such as the retrieval of knowledge. A lexicon plays a significant role in any kind of language processing application. There are many advantages to a corpus-based approach, rather than using traditional approaches. Based on a text obtained from a corpus of 10 million words [23], the lexicon developed for Sinhala has been based on the research paper that authors studied on this approach. The program should gradually evolve according to the revised lexicon to propose various ways to correct a grammatical error. Not only that, but it's also a neural network system that is another way to do this job. There are two different methods of being introduced—classification of text based on class, and classification of text by mathematical model. Sentences are separated into courses to train in class- based text classification. In that case, more set of data with more classes is needed.

This method of classification of texts has already been carried out for English. It educated as a greeting, goodbye, and sandwich kind of categories. Set of data implemented as a wise category. Then he tried a new word that matched one of the current groups. Since this is a mathematical model, the system converges between 0-100 between 14 ends with an intermediate value. If the phrase is accurate, the output percentage value should be at least greater than 50%. If the sentence performance value is wrong, the percentage value should be less than 50%. The accuracy of that model can also be achieved by considering the values of the output. The suggested technique will go much more accurately than the current grammar testing program.

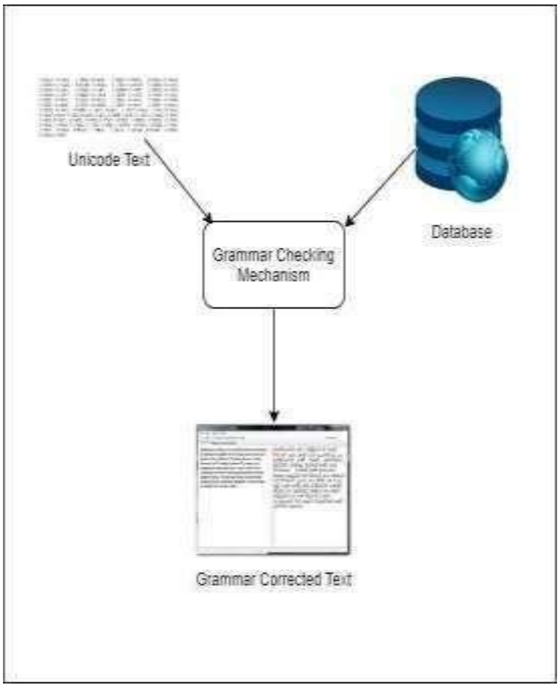

Figure 5: High Level Architecture Diagram

Besides, this computational grammar model can be used to develop various types of computer applications based on the Sinhala language such as spell and grammar checkers, word generator, etc. English to Sinhala machine translation system can be evaluated via stranded metrics as further work of this project. The morphological Sinhala generator accesses the Sinhala dictionary and generates suitable Sinhala word forms. This morphological generator was designed for the generation of appropriate Sinhala words through the Sinhala grammar rules. It needs appropriate Sinhala base word, number, case, and noun form (direct or indirect) to generate a Sinhala Noun. The Sinhala verb requires tense, person, and number of the appropriate Sinhala base-word. The adjective to the Sinhala requires the type and the base word. The type of Adjective can be positive, comparative, or superlative [24]. The other word type is not a part of the word conjugation. Those words are therefore stored in the Sinhala dictionary. These words are directly read by the Sinhala morphological generator from the Sinhala dictionary.

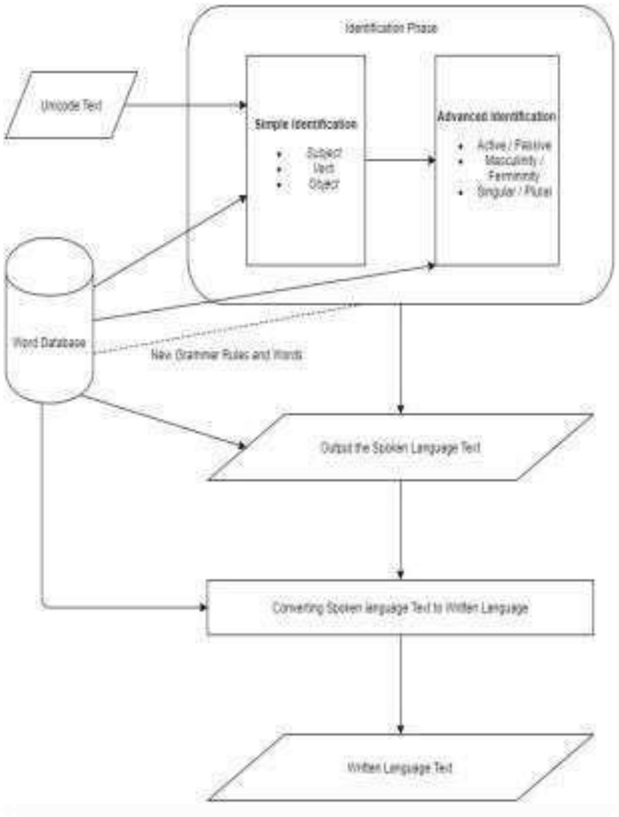

Figure 5.1: Blue print of Grammar Correction Mechanism

V. RESULTS AND DISCUSSION

This paper provides an overview of the current state-of-the-art voice recognition systems and highlights several challenges and problems that will have to be addressed in future years. The diarization of speakers is not yet sufficiently mature to allow methods to be easily transported across different domains where minor differences in meeting data (recorded at the same sites) lead to wide variations in results. Meanwhile, more massive datasets need to be compiled to make results more meaningful and to make systems more robust to unseen variations. To process these data, systems would have to become more potent with growing data set sizes. Perhaps the most considerable single difficulty is the management of simultaneous words, to be attributed to multiple speakers. As a reasonably embryonic culture, there are, therefore, excellent prospects for significant advances and substantial improvements to the rather ad hoc and heuristic solutions that currently dominate the area, at least in comparison with the more developed fields of speech and speech recognition. This area seems much more comprehensive and lighter than the present, as more and more people understand the utility of audio methods for many activities historically considered to be solvable solely in the visual domain. Recognition of speakers is one of the fundamental problems that underlie virtually every activity involving acoustics and the presence of more than one person. The grammar was tested to Sinhala Machine Translation System via the English. This system was developed as a rule-based framework and the translation method of English Morphological Analyzer, English Parser, Sinhala Base Word Translator, Sinhala Morphological Generator, and Sinhala Sentence Composer. As the theoretical basics of translation, the concept of conjugation is used. This translates English sentences with basic, complex subjects and objects with tense patterns most commonly used. This method operates for a small domain in addition to the above, and it relies on the word availability of the dictionaries and the accuracy of the papers, analyzers, and generator. This research paper is a collective effort of the generation of the Sinhala speech recognition system to a decoded parameter. The researchers found the main aim of assisting the people who fight Dyslexia successfully through the development of this speech recognition system. Moreover, this project catered to the significant social development of producing accurate data and information to enhance communication.

VI. FUTURE WORK

Future work includes adding support for multiple languages, increasing the accuracy of transcripts and adding support for videos as well.

VII. ACKNOWLEDGEMENTS

Our university, the Sri Lankan Institute of Information Technology (SLIIT), has supported this research in the generation of more practical approach. Authors would like to thank SLIIT for providing this research with insight and expertise. We also wish to express our gratitude to individual anonymous reviewers for their insights.

VIII. REFERENCES

[1] Ilker Bayram, "Employing phase information for audio denoising," IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[2] Gautam Bhattacharya and Philippe Depalle, "Sparse Denoising of Audio by Greedy TimeFrequency Shrinkage," 2014 IEEE International Conference on Acoutics, Speech and Signal Processing (ICASSP), 2014.

[3] Li Zhang, Xiaomei Chen, Bo Zhong, Longbiao He, Huan Xu, and Ping Yang "Objective Evaluation System for Noise Reduction Performance of Hearing Aids" 2015 IEEE International Conference on Mechatronics and Automation August 2-5, Beijing, China.

[4] Y. Ramon, "Speaker Diarization with Kaldi," *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/speaker-diarization-with-kaldi-e30301b05cc8.

[5] S. Meignier, J.-F. Bonastre, S. Igounet, E-HMM approach for learning and adapting sound models for speaker indexing, in 2001: a Speaker Odyssey. The Speaker Recognition Workshop, Chania, Crete, 2001, pp. 175180.

[6] J. Ajmera, C. Wooters, A robust speaker clustering algorithm, in Automatic Speech Recognition and Understanding, IEEE, ASRU 2003, St. Thomas, U.S. Virgin Islands, 2003, pp. 411416.

[7] "Real-time Translation of Discrete Sinhala Speech to Unicode Text", https://www.researchgate.net/, 2020. [Online]. Available: https://www.researchgate.net/publication/281434759 _Real_time_Translation_of_Discrete_Sin hala_Speech_to_Unicode_Text.

[8] "A Real-Time Speech to Text Conversion Technique for the Bengali Language," 2020. [Online]. Available: https://www.researchgate.net/publication/327817644.A_ Real_Time_Speech_to_Text_Conver sion_Technique_for_Bengali_Language.

[9] "Audio to Text Converter | Kukarella", Kukarella.com, 2020. [Online].Available: https://www.kukarella.com/audio-to-text-converter.

[10] Hettige.B, Karunananda A.S, "Project මමමමමමමු," University of Moratuwa, 2011 [Online]. Available: https://ieeexplore.ieee.org/document/6075022?reload=true&arnumber=6075022

[11] Lahiru Abeyrathne, Sarangi Edirisinghe, Rumesh Premachandra, Apsaari Warsha, Nalaka De Silva, S. Thelijjagoda, "Spell and Grammar Checking Tool for Sinhalese මමමමමසම මමම - මමමම මමමමමමම මමමමමමමු," Department of Information Technology, Department of Information Systems Engineering, Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka. Department of Sinhala, Faculty of Arts, University of Colombo. [Online]. Available: http://multidisciplinaryjournal.globalacademicresearchinstitute.com/images/engineering/W.A.%20Lahiru%20Dasun%20Abeyrathne.pdf.

[12] Herman Kamper, Aren Jansen, and Sharon Goldwater. 2016b. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. IEEE/ACM Trans. Audio, Speech, Language Process., 24(4):669–679.

[13] J. VanderPlas, "Understanding the FFT Algorithm | Pythonic Perambulations," Jakevdp.github.io, 2020. [Online]. Available: https://jakevdp.github.io/blog/2013/08/28/understanding-the-fft/.

[14] "What is unsupervised learning? - Definition from WhatIs.com", WhatIs.com, 2020. [Online]. Available: https://whatis.techtarget.com/definition/unsupervised-learning.

[15] Alex S. Park and James Glass. 2008. Unsupervised pattern discovery in speech. IEEE Trans. Audio, Speech, Language Process., 16(1):186–197.

[16] Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). *Speaker Diarization: A Review of Recent Research. IEEE Transactions on Audio, Speech, and Language Processing, 20(2), 356–370.*

[17] "Feature Extraction", *DeepAI*, 2020. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/feature-extraction.

[18] M. Jumelle, "Speaker Clustering", *https://arxiv.org/*, 2020. [Online]. Available: https://arxiv.org/abs/1803.08276.

[19] Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., … Reynolds, D. A. (2004). *A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Advances in Signal Processing, 2004(4).*

[20] "Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)", Ijstr.org, 2020. [Online]. Available: http://www.ijstr.org/final-print/june2015/Speech-totext-Conversion-stt-System-Using-Hidden-Markov-Model-hmm.pdf.

[21] P. Taylor, A. Black, and R. Caley, "The architecture of the festival speech synthesis system," Cs.cmu.edu, 1999. [Online]. Available: https://www.cs.cmu.edu/~awb/papers/ESCA98_arch.pdf

[22] Wikipedia, the free encyclopedia "Natural Language Processing" last edited in January 2020. [Online].Available: https://en.wikipedia.org/wiki/Natural_language_processing

[23] R. Jensen, J. Gatrell, J. Boulton, and B. Harper, "Using Remote Sensing and Geographic Information Systems to Study Urban Quality of Life and Urban Forest Amenities," Ecologyandsociety.org, 2004.[Online]. Available: https://www.ecologyandsociety.org/vol9/iss5/art5/print.pdf.

[24] Hettige, B., & Karunananda, A. S. (2011). *Computational model of grammar for English to Sinhala Machine Translation. 2011 International Conference on Advances in ICT for Emerging Regions (ICTer).*