ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse02952020.pdf

https://doi.org/10.30534/ijatcse/2020/02952020

# A Game Application to assist Speech Language Pathologists in the Assessment of Children with Speech Disorders

**Dhanya Sasikumar[1], Saumya Verma[2], K. Sornalakshmi[3]***

[1]Department of Information Technology,
College of Engineering and Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology,
SRM Nagar, Kattankulathur, 603203, Kanchipuram, Chennai , TN, India, dhanya.official997@gmail.com

[2]Department of Information Technology,
College of Engineering and Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology,
SRM Nagar, Kattankulathur, 603203, Kanchipuram, Chennai , TN, India, saumyav65@gmail.com

[3]Department of Information Technology,
College of Engineering and Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology,
SRM Nagar, Kattankulathur, 603203, Kanchipuram, Chennai, TN, India, sornalak@srmist.edu.in

## ABSTRACT

Speech Disorders in children is the fifth highest occurring disability in India. Speech Disorders have a prevalence of 8 to 9 percent in children. Thus early diagnosis and treatment can increase the recovery rate among children in the age group of 3-12 years. Due to the shortage in the number of Speech Language Pathologists all over world, it becomes impossible for everyone to seek treatment. This brought forth a need for alternative modes of Speech Therapy. Computer-Based Speech Therapy is the most preferred alternative mode of speech therapy due to its availability and effectiveness. But most Computer Based applications are very expensive and cannot be afforded by the common man. In this paper we introduce a game application which assesses the speech of a child by analyzing the voice activity and the mispronounced phonemes in the speech. This application is aimed to provide assistance to Speech Language Pathologists.

**Key words:** Automatic Speech Recognition, Feature Extraction, Mel Frequency Cepstral Coefficients, Preprocessing, Phoneme Detection, Voice Activity Detection

## 1. INTRODUCTION

Speech disorders are basically communication disorders that disrupt normal speech. Speech disorders are the fifth highest in our country with a prevalence of 8 to 9% in children.

People suffering from speech disorders are often subjected to social, physical and mental challenges. Therefore, it becomes a necessity to aid children with speech impediment as early as possible. Early detection will lead to early treatment thus improving the chances of recovery.

There are various levels of classification that is used to determine the types and the severity of Speech disorders thereby helping to find proper therapy and treatment for the patients. There are three main levels used to determine a pedagogy for therapy. 1) Phonemic or phonetic sounds produced. Phonemic sounds are easily produced and are used to distinguish one word from another whereas phonetic sounds are single letter pronunciations produced only when required. 2) Stimulation of sound. Either the sound could be easily stimulated or stimulation only occurs after probing or demonstration. 3) Voluntary production of sound. Either there is no voluntary production of sound or there is no production of sound at all. On the basis of these levels we SLPs can gauge the severity of any type of speech disorder. There are various types of speech disorders.

- Apraxia of Speech can be a result of a stroke or some progressive illness. There is inconsistency in the production of speech and it involves the re-arrangement of the sounds in a word. There usually is some difficulty in producing words but some common phrases are sometimes effortlessly spoken.
- Cluttering is a condition wherein the patient speaks too rapidly which makes the understanding of the speech difficult.
- Developmental Verbal Dyspraxia or Childhood Apraxia of Speech is a neurological pediatric speech sound and pronunciation.
- Dysarthria is caused due to damage to the nerves or muscles responsible for speech. It can also be caused by Parkinson’s disease, strokes and Cerebral Palsy.
- Dysprosody is a neurological disorder due to which a person may speak fluently but the intonation and the rhythm of words, intensity of the sounds and the timings of the utterances is disrupted.
- Muteness is when a person has total inability to speak.
- Phonetic disorder is when there is difficulty in learning how to produce sounds.

- Phonemic disorder is when is when there is difficulty in learning the distinction in sounds.
- Stuttering is a disorder that involves a perturbed flow of speech.

There are various technologies available that can enable people with speech impediment to speak with very little effort. Existing software tools do provide online Speech Therapy and especially tailored programs that enable children to learn faster. However, SLPs are under immense pressure to provide a personalized treatment to each and every patient. More so because even with a considerable population suffering from various kinds of speech disorders there is a lack of SLPs in our country. Thus, there is a need for a virtual assistant for SLPs which is accessible at anytime from anywhere.

In order to achieve this we have created a virtual assistant in the form of a desktop gaming application which consists of various levels. Each level presents a word which is to be spoken by the child and will be recorded for assessment. This speech input is passed through two modules 1) Phoneme Detector 2) Voice Activity Detector. Phoneme Detector lists out the mispronounced phonemes. The Voice Activity Detector detects the speech and non-speech which the SLP can use to identify if groping errors are present in the speech. This application provides a second opinion to the SLP which saves him/her some time thus increasing the reach of SLP to a larger group of patients.

## 2. BACKGROUND AND RELATED WORKS

### 2.1 Background of Speech Therapy

Speech Pathology began in the early 1920s when there was a need for it. The "American Academy of Speech Correction" was established in 1926. It developed more and more over the next twenty years. During this period, World War II was going on and a lot of soldiers returned home with injuries to their brain which caused problems in their speech. They were treated by Speech Pathologists.

Speech therapy witnessed advancements over the years. The number of Speech Institutions and the number of conditions being treated increased manifold as was the research in this field. However, with increased number of patients the corresponding number of Speech Pathologists increased rather slowly. This brought forth the need for alternative modes of Speech Therapy. Computer-aided Speech Therapy is an effective means of remote treatment of Speech Disorders. It assists the children with speech disorder to get effective treatment at the comfort of their own home while being remotely assessed by SLP. Speech Therapy has come a long way in the 21$^{st}$ Century. Computer-based Speech Therapy or CBSTs are a viable option for those who do not have access to SLPs. Various Speech Therapy applications aim at providing treatment to patients in the form of exercises and games. The exercises are designed by Speech Pathologists and have varying levels of difficulty. The child is evaluated based on his/her performance in the exercises and appropriate feedback is given.

The software tools available for speech therapy are beneficial for children but they are expensive and sometimes complicated. Moreover, none of the applications reduce the work of an SLP or provide him with a second opinion.

### 2.2 Related Works

Since the early 1920s, a lot of work has been done in this field resulting in many Computer based Speech Therapy Software. Many of these software were in the form of a game to make the therapy sessions fun and engaging. However, Beena Ahmed et al [1] found that existing games lacked an integrated feedback mechanism. Hence, they created speech controlled game application and evaluated the performance of the application based on the experience of the children and the SLPs. They also included a feedback mechanism. Overall, the children and SLPs were satisfied with the performance of the app barring a few features.

Nuffield Dyspraxia Program is an intervention programme for children in developmental age of 3-7 years with severe speech sound disorders. It is a complete treatment package which includes assessment and therapy procedures and a vast source of pictorial materials to provide pedagogy for treatment of children. Avinash Parnandi et al [2] developed a mobile application based on Nuffield Dyspraxia Program. This application has a multi-tier client-server architecture and provides remote administration of speech therapy. The aim of this application was to decrease the dependence of children on SLPs as there is a lack of professionals to provide proper treatment and physical availability of SLPs is not possible in remote areas.

There are many assessment software tools available for SLPs but most of them use English Language making it difficult for Filipino patients to follow. Thus Marilou N. Jamis et al [3] developed a mobile application called Speak App to help patients practice in Filipino language.

Alternative and Augmentative Communication is a culmination of methods that supplement spoken or written language for those with impairments in the production or comprehension of spoken or written language. Erh-Hsuan Wang et al [4] noticed that the existing AAC technologies have limitations due to which best results are not obtained. Thus a better AAC android application was created using VoiceText Embedded SDK. However, the usability score of the application is low and it does not provide assistance to the SLPs.

Achieving human like performance in machines to identify speech is difficult. The error rate for machines is higher than that for humans because humans have more adaptability to learn new words and operate in noisy conditions. Also, humans can accurately recognize non-sense words as they are less reliant on context. To overcome this shortcoming Latanya Sweney and Patrick Thompson built a system called the BeBe system the purpose of which was to detect phonemes reliably and consistently. They came up with separate detection algorithms for every phoneme. However, the accuracy for this system was only 83%.

The BeBe system does identify various phonemes but it does not use artificial intelligence hence it is a heavy application that too with low accuracy [5]. It will be safe to conclude that BeBe system will not go a long way to assist SLPs. Other aforementioned applications fail to provide assistance to the SLP. In contrast to our system, which instead of replacing the SLP makes their job easier. The speech-processing capabilities provide a second opinion to the SLP reducing the time given to each patient and allowing them to reach a larger number of patients.

## 3. SYSTEM OVERVIEW

The Proposed System is a Game Application made using Tkinter Graphical User Interface (GUI). The User (patient) as well as the Speech Language Pathologist (SLP) can interact using this Interface. The application starts with the user logging in or registering to make an account if he does not already have one. Once the user logs in, the game starts. The game has multiple levels of varying difficulty. Every level shows a word and an image representation of the word. The child/user has to speak the word through the microphone. The speech input is passed on for pre-processing. Figure 1 shows the proposed System Architecture.

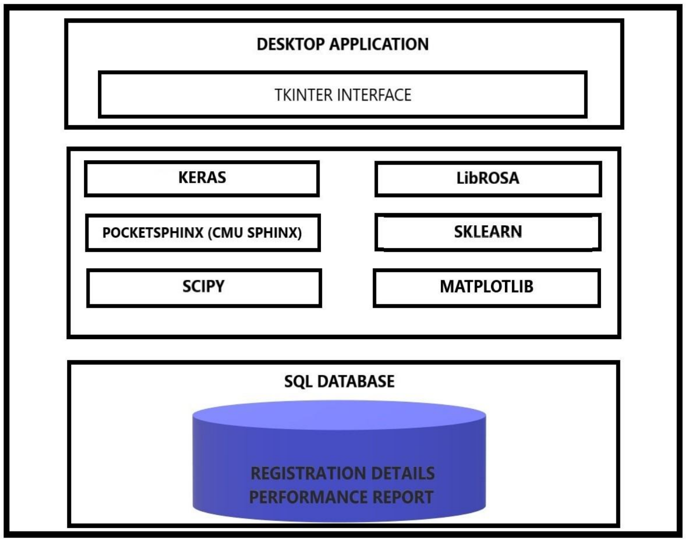


**Figure 1:** System Architecture

### 3.1 Preprocessing

The input speech is stored separately which is in '.wav' format. The wav files have an underlying pulse code modulation data which contains the amplitude of the continuous time signal. However, the audio signal received as input from the patient cannot be directly used. It needs to be converted into digital signal that can be crunched down into numerical arrays which can be used by upcoming processes. This conversion is performed by the pre-processor by using the sampling technique. Sampling can be defined as a method that converts continuous time signal into a discrete one. A sampler is built to extract samples from continuous time signal. This sequence of samples when passed through a reconstruction filter gives the original signal. The number of samples stored per second is known as the sampling rate which by default is set 44.1kHz.

For speech recognition, if the audio is found to be stereo i.e if the number of channels is equal to two then the audio signal is transformed into mono i.e channel is equal to one. Now that a mono channel audio signal has been obtained, removal unwanted elements or noise is the next step [6]. For this certain calculation need to be performed on the signal. The further calculations on the signal can be done either in time domain or in frequency domain. However, some operations are easier to perform in the frequency domain. Thus a transform is used to convert time domain into frequency domain. Usually a Fast Fourier Transform is used (FFT). A filter can now be easily used to remove noise and DC offsets.

### 3.2 Feature Extraction

After Pre-processing, the next stage we have is Feature Extraction. Feature Extraction is the process of extracting the essential features from a particular set of data so that it can be used in our Speech Recognizer, Phoneme Detector and Voice Activity Detector. In this process, huge complex sets of data is reduced so that it becomes easy to process them. It helps in identifying important data while leaving out the redundant and irrelevant ones. Using this method, we can either combine various features or select a few essential ones. It can be used find and analyse the relationship between the data. The features are going to be extracted using the LibROSA library in Python.

Zero Crossing Rate gives us the rate at which changes in sign occur through a signal. It gives the rate of change of a signal from negative to zero and then to positive and from positive to zero and then to negative [7]. Other features that are extracted include Spectral Centroid, Spectral Rolloff and MFCC.

Mel-Frequency Cepstral Coefficients are one of the most important features extracted from audio signals. The MFCCs of a signal are features that gives us the shape of the spectral envelope. It is usually is small sets of 10 to 20. *Mel Scale* is a scale that is used to scale frequencies so that it is very close to the human-perceivable form [8]. The Mel Scale captures the differences in the perceived distances of sound. A Filter bank is used to convert the frequencies to Mel Scale and are also called Mel Filters. Using the shape of a person's vocal tract you can determine any sound generated by them. The representation of any sound can be accurate if this shape is obtained correctly. MFCCs are the coefficients of the Mel-frequency cepstrum and accurately represents the envelope of the time spectrum of the audio signal which represents the vocal tract.

### 3.3 Automatic Speech Recognition (ASR)

The next stage is Automatic Speech Recognition. The speech input provided by the child is analyzed by the ASR to check if he/she has spoken the word correctly or not. If the word was spoken correctly, the game will move to the next word else further analysis of the speech input is made to check where the child is going wrong. Automatic Speech Recognition is a method that converts speech to text using the features extracted in the previous stage and Neural Networks [9].

Building an ASR involves various steps. Firstly, we resample the signal to 8000 Hz since it is the most common speech frequency. After that, we extract all the .wav files from the dataset and store it in a list. Resample the audio signals and remove all the speech commands that are as short as 1 second. Label Encoder helps convert the output labels into integer encoded labels. These labels are further converted to a one-hot vector since we are dealing with multi classification problem. We split the dataset into training and validation set [10]. The Speech-To-Text model is created using Convolutional Neural Networks (CNN). CNNS use Convolutions and have Pooling Layers in its network structure. We reshape the 2D input array to a 3D array because conv1D takes 3D array as input. Conv1D is a 1 dimensional convolutional neural network where we have the kernel sliding along one dimension and has spatial properties. Conv1D is used on time-series data. Since, text and audio can be represented as a time series data, we use conv1D on them. The Speech-to-Text model is created using the Keras module in Python. The ASR is responsible for correctly recognizing the words spoken by the child. If the child speaks the word correctly, he/she gets 10 points and the game proceeds to the next level. If the child does not speak the word correctly, the speech input is further analysed to understand where the child is going wrong thereby helping the Speech Pathologist understand how to proceed with the treatment. The exercises and the levels are modelled based on the evaluation made on the speech input.

The evaluation is done using two methods – Phoneme Detection and Voice Activity Detection.

### 3.4 Phoneme Detection

The next stage of the game is Phoneme Detection. In case the child is unable to speak the given word correctly, the ASR would not recognize it and hence it passes to the evaluation phase. The first step in the evaluation phase is Phoneme Detection. A phoneme is the smallest component of sound that divides a word into individual parts. For example, the phonemes of the word SEVEN are S, EH, V, AH and N. A spoken word can be broken down into phonemes to know exactly which part of the word the child is having struggle pronouncing.

Phoneme Detection algorithms need require high accuracy to avoid misdiagnosis. To overcome this, we use the phoneme recognition tool provided by CMU Sphinx**,** the features of which are available in the pocketsphinx library [11]. It has multiple inbuilt phonetic language models. A decoder is used to record the input utterances and give a hypothesis in the form of a list of phonemes.

Phoneme recognition allows us to correctly identify the phonemes that are wrongly being pronounced by the child. Phonological speech disorders are very prevalent in small children and can affect the performance of the child in general. Hence, it becomes essential to improve the child's speaking abilities. This can achieved by modelling the exercises in a way such that the game provides more words that contain the phonemes the child is unable to speak properly. Therefore, Phoneme Recognition plays an important part in analyzing the child's need.

### 3.5 Voice Activity Detection (VAD)

Voice Activity Detection is the process of determining whether a signal contains speech or not. Thus, it can be considered as a binary decision maker as the desired output target is either 0 for absence of speech or 1 for presence of speech. A subtask of this step is to determine the Speech Presence Probability (SPP). It is expressed in the range of 0 to 1. The output of the speech presence probability estimator is thresholded to give the voice activity classification [12]. Speech provides additional energy to the signal and the high-energy regions of the speech input are most likely to be speech. Therefore, we set a threshold value for energy such that when the energy of the signal is above the threshold, it is indication of speech activity. The performance of VAD mainly depends on the choice of the threshold. Appropriate value of threshold is generally chosen by trial and error method.

To improve the performance of Voice Activity Detector, more features of the speech signal were used for defining the speech regions. Linear predictors define the shape of audio signals with more efficiency. A low modelling error implies that the signal is likely a speech signal. To make a stronger case, we check if there is a significant fundamental frequency within the range of 80 and 450 Hz, then the signal is a speech signal. Otherwise, the signal is classified as a non-speech signal. The Scipy library in python has functions that help us in building a Voice Activity Detector.

## 4. EXPERIMENTS AND RESULTS

The Game Application was tested using the TORGO Dataset. It contains audio data of two categories – People with Dysarthria and people without Dysarthria. Each category has both male and female subjects. The dataset includes people with Cerebral Palsy and people with Amyotrophic Lateral Sclerosis.

The Speech-To-Text model used in the ASR had an accuracy of 91%. The accuracy was calculated while building the model by continuously comparing and monitoring the accuracy after every epoch. The CMU

Sphinx Phoneme Recognition tool gives around 84% accuracy. The VAD algorithm was tested using the TORGO Dataset and it accurately labelled speech and nonspeech regions.

The first window that opens on executing the application is shown in Figure 2.

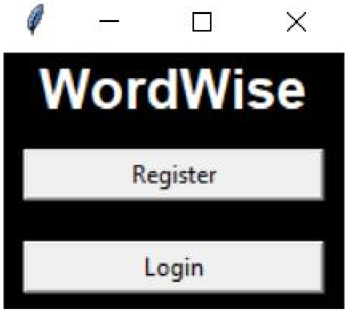
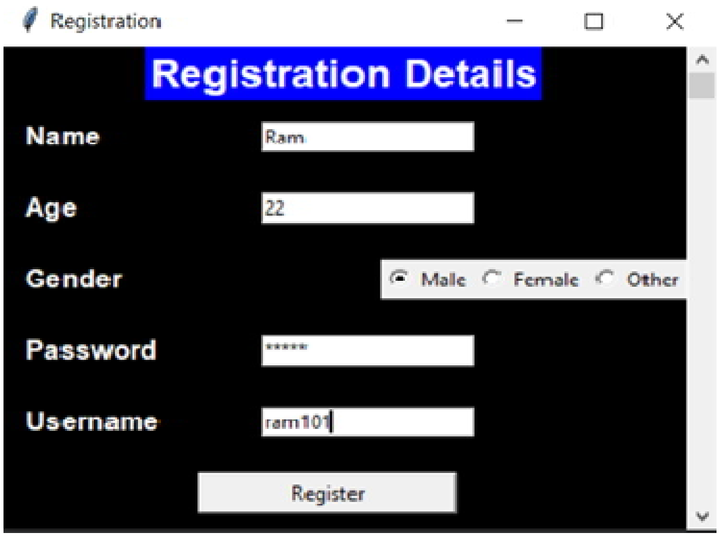


**Figure 2:** First Window of the Application

Figure 3 and Figure 4 show the Registration and Login windows respectively.



Name: Ram

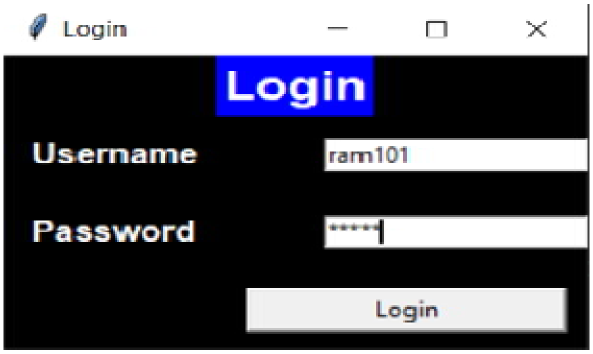
Age: 22

Gender: ● Male ○ Female ○ Other

Password: *****

Username: ram101

**Figure 3:** Registration Window



Username: ram101

Password: *****

**Figure 4:** Login Window

The game screen that appears after the user logs in is shown in Figure 5. The game has multiple levels of varying difficulty. Every level shows a word and its image representation. The child (user) is supposed to click on the speak button and say the word through the microphone.

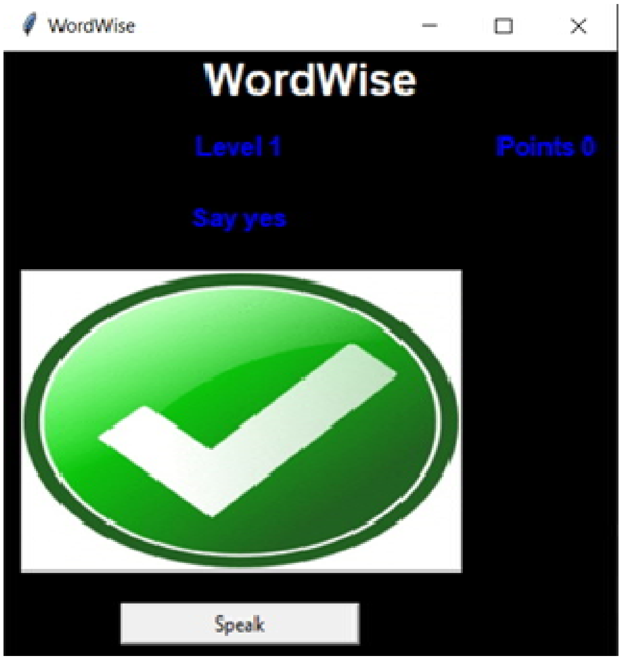
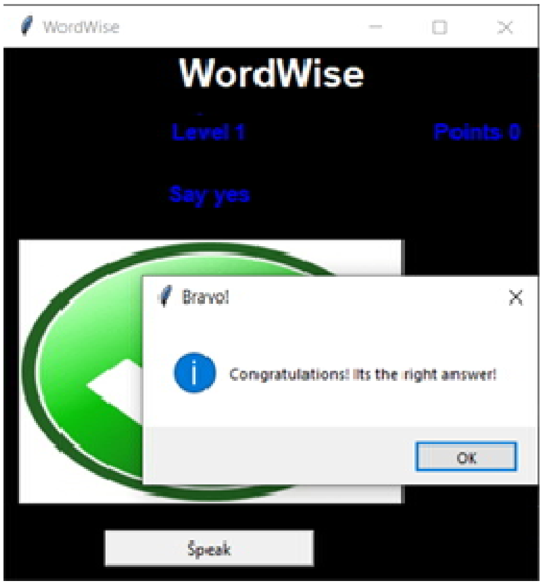


**Figure 5:** Game Screen

If the user speaks the word correctly, he/she gets 10 points and the game moves to the next level as shown in Figure 7. A congratulatory message is displayed after every correct answer to boost the morale of the user and is shown in Figure 6.



Congratulations! Its the right answer!

**Figure 6:** Message shown after correct input

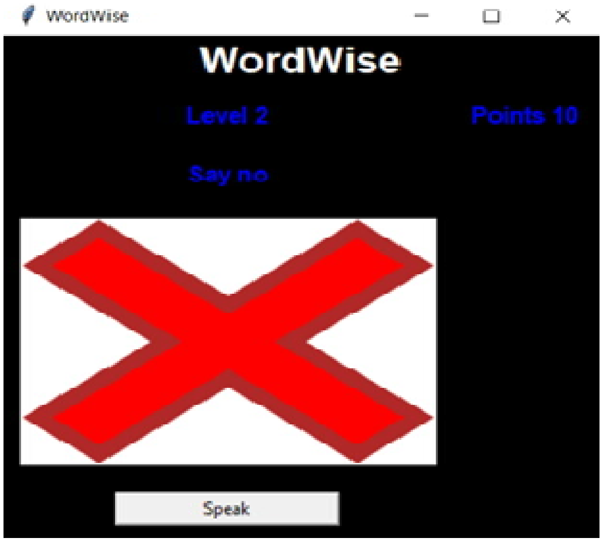


**Figure 7:** User gets 10 points and game moves to next level.

Finally, after all the levels are over, a report can be generated. Figure 8 shows the Report window.

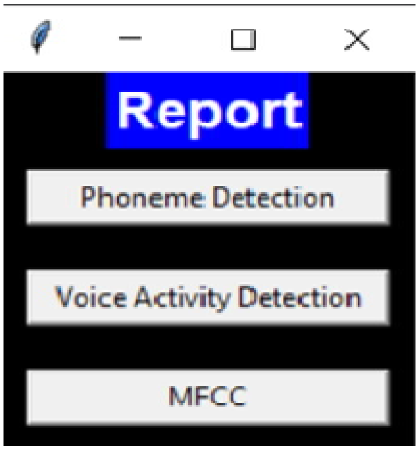
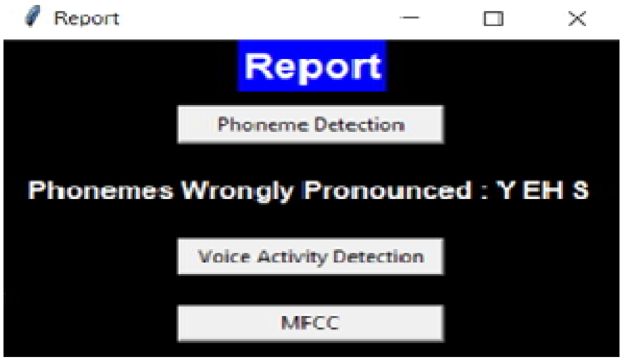


**Figure 8:** Report Window

The Phoneme Detection window gives us a list of mispronounced phonemes as shown in Figure 9.



Phonemes Wrongly Pronounced : Y EH S

**Figure 9:** Mispronounced Phonemes

The MFCC window gives us a graph that compares Original Speech and Input Speech as shown in Figure 10.

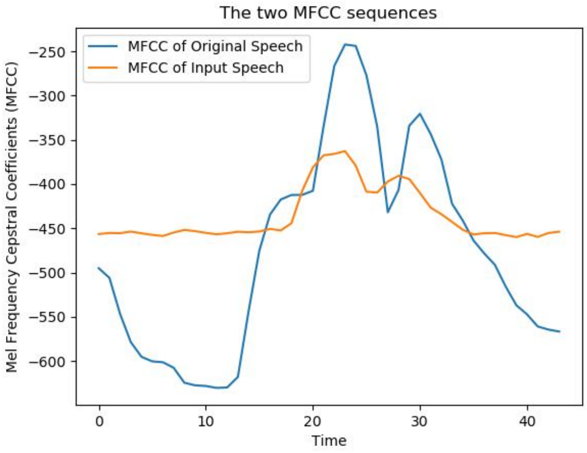
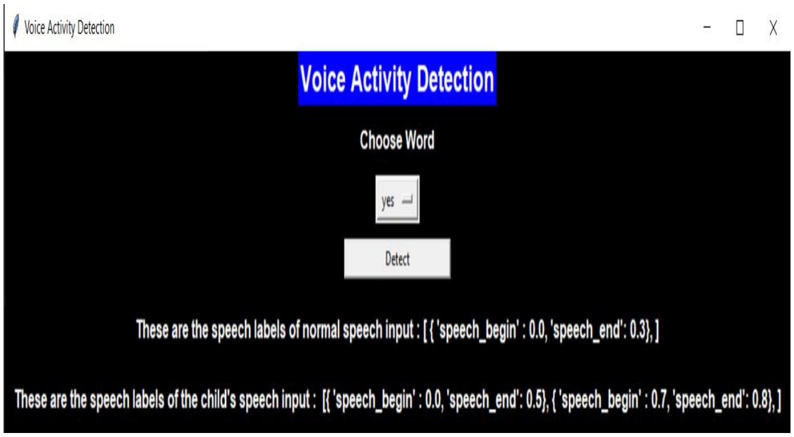


**Figure 10:** MFCC values of Original Speech vs. Input Speech plotted against Time.

Finally, the Voice Activity Detection window gives us the Speech Labels of Normal Speech and Input Speech as shown in Figure 11.



```
These are the speech labels of normal speech input : [ { 'speech_begin' : 0.0, 'speech_end': 0.3}, ]

These are the speech labels of the child's speech input :  [{ 'speech_begin' : 0.0, 'speech_end': 0.5}, { 'speech_begin' : 0.7, 'speech_end': 0.8}, ]
```

**Figure 11:** VAD displaying the speech labels of normal and input speech.

## 5. CONCLUSION

Speech and communication skills are necessary for humans to develop relations and get their day-to-day affairs in order. However, for many children as well as adults speech is an impediment. Speech Disorders have a prevalence of 8 to 9 percent in children. There is a shortage in the number of Speech-Language Pathologists around the world. Sometimes SLPs can feel overwhelmed by the workload given that each patient has specific requirements. SLPs need assistance to fast forward the assessment process so that the intervention can be provided as soon as possible.

Moreover, most Computer Based Speech Therapy tools are expensive and/or has geographic limitations. For this purpose we designed a game application called WordWise. The WordWise app can be used by Speech Language Pathologists to help assess Children with Speech Disorders. It is designed as a game so that it is fun and engaging for the kids. It is inexpensive, user-friendly and can be used remotely from the comfort of your home. When the application was tested using the TORGO dataset the results seemed to provide accurate assessment of the patients with dysarthria. The SLP can view the reports of the VAD module, Phoneme Detection and MFCCs and design suitable intervention pedagogy for his/her patient.

## 6. FUTURE ENHANCEMENTS

This application has scope for improvement. Starting with the UI it can be made user-friendly and attractive for children. The ASR can be trained better to recognize a wide range of words. For phoneme detection module, it is still possible to achieve higher accuracy. The VAD module can be modified to give the final report itself thus further reducing the workload of the SLP. Finally, this application can be converted into a mobile application so that it can reach a wider population of SLPs.

## 7. REFERENCES

[1] Beena Ahmed, Penelope Monroe, Adam Hair, Check Tien Tan, Ricardo, Gutierrez-Osuna and Kirrie J. Ballard. **Speech-Driven Mobile Games for Speech Therapy: User experiences and feasibility,** *International Journal of Speech-Language Pathology*, Vol 20, Issue 6, 2018.

[2] Avinash Parnandi, Virendra Karappa and Youngpo Son, **Architecture of an Automated Therapy Tool for Childhood Apraxia of Speech,** *ASSETS 2013: Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, Article No. 5, pg. 1–8, October 2013.

[3] Marilou N. Jamis, Emeliza R. Yabut and Rosauro E. Manuel, **Speak App: A Development of Mobile Application Guide for Filipino People with motor Speech Disorder,** *TENCON 2018 – 2018 IEEE Region 10 Conference*, 28-31 Oct. 2018.

[4] Erh-Hsuan Wang, Leming Zhou and Szu-Han Kay Chen, **Development and Evaluation of a Mobile AAC: A Virtual Therapist and Speech Assistant for People with Communication Disabilities,** *Disability and Rehabilitation: Assistive Technology* , Vol. 13, Issue 8, 2018.

[5] Latanya Sweney and Patrick Thompson, **Speech Perception Using Real-time Phoneme Detection: The BeBe System.**

[6] Nishan Singh and Dr. Vijay Laxmi, **Audio Noise Reduction from Audio Signals and Speech Signals,** *International Journal of Computer Science Trends and Technology (IJCST)*, Vol. 2, Issue 5, Sep-Oct 2014.

[7] https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d

[8] https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

[9] D. Nagajyothi and P. Siddaiah, **Speech Recognition Using Convolutional Neural Networks,** *International Journal of Engineering and Technology*, Vol. 7, (4.6), pg. 133-137, 2018.

[10] https://www.analyticsvidhya.com/blog/2019/07/learn-build-first-speech-to-text-model-python/

[11] https://cmusphinx.github.io/wiki/phonemerecognition/

[12] Tom Bäckström, **Voice Activity Detection - Speech Processing,** *Speech Coding*, pp. 185-203, 2017.