Automatic Speech Recognition System to Analyze Autism Spectrum Disorder in Young Children

M.Abinath[1], S.M.Afaal[2], P.Anojan[3], R.Vithurshan[4], Prof. Koliya Pulasinga[5] and Wishalya Tissera[6]
[1]Department of Information Technology, Sri Lanka Institute of Information Technology, SRI LANKA
[2]Department of Information Technology, Sri Lanka Institute of Information Technology, SRI LANKA
[3]Department of Information Technology, Sri Lanka Institute of Information Technology, SRI LANKA
[4]Department of Information Technology, Sri Lanka Institute of Information Technology, SRI LANKA
[5]Department of Information Technology, Sri Lanka Institute of Information Technology, SRI LANKA
[6]Department of Information Technology, Sri Lanka Institute of Information Technology, SRI LANKA

[1]Corresponding Author: it19242828@my.sliit.lk

ABSTRACT

It's possible to learn things about a person just by listening to their voice. When trying to construct an abstract concept of a speaker, it is essential to extract significant features from audio signals that are modulation-insensitive. This research assessed how individuals with autism spectrum disorder (ASD) recognize and recall voice identity. Autism spectrum disorder is the abbreviation for autism spectrum disorder. Both the ASD group and the control group performed equally well in a task in which they were asked to choose the name of a newly-learned speaker based on his or her voice. However, the ASD group outperformed the control group in a subsequent familiarity test in which they were asked to differentiate between previously trained voices and untrained voices. Persons with ASD classified voices numerically according to the exact acoustic characteristics, whereas non - autistic individuals classified voices qualitatively depending on the acoustic patterns associated to the speakers' physical and psychological traits. Child vocalizations show potential as an objective marker of developmental problems like Autism. In typical detection systems, hand-crafted acoustic features are input into a discriminative classifier, but its accuracy and resilience are limited by the number of its training data. This research addresses using CNN-learned feature representations to classify children's speech with developmental problems. On the Child Pathological and Emotional Speech database, we compare several acoustic feature sets. CNN-based approaches perform comparably to conventional paradigms in terms of unweighted average recall.

***Keywords*—** AI, ASD, CNN

I. INTRODUCTION

Speaker diarization is a task that involves labeling audio or video recordings with classes that correlate to speaker identification. In other words, it is a task that attempts to determine "who spoke when." Early on in the field of voice recognition on multispeaker audio recordings, speaker diarization techniques were developed to enable speaker adaptive processing. These algorithms have also earned their own usefulness over time as a stand-alone application, which enables them to give speaker-specific meta information for use in subsequent tasks such as audio retrieval. More recently, with the advent of deep learning technology, which has driven revolutionary changes in research and practices across speech application domains, rapid advancements have been made for speaker diarization [1]. These changes have been brought about as a direct result of deep learning's influence. Autism Spectrum Disorder (ASD) is a developmental condition that is characterized by difficulty in social communication, social interactions, and repetitive activities. Some of these challenges are reflected in the speech patterns of children with autism spectrum disorder (ASD) who are talkative. The development of algorithms that are able to extract and quantify speech elements that are unique to children with ASD is, as a result, incredibly important for analyzing the starting state of each kid as well as their growth throughout the course of time. Speaker diarization is a crucial part of such algorithms, particularly in the loud clinical situations in which ASD children are diagnosed [2].

Speech recognition technologies are becoming increasingly important in the lives of all people nowadays. It is a piece of software that gives users the ability to communicate with their mobile devices by using only their voices. The audio of a speech is dissected into its component sound waveforms by speech recognition software, which then conducts an analysis of each sound form, and applies a number of different algorithms to find the word that is the best fit for the sound in that language [3], and transcribes the sounds into text. Instead of using various instruments to operate an electronic device, such as keystrokes, buttons, keyboards, etc., the user can instead operate the device only through spoken words thanks to a type of technology known as speech recognition. The words and phrases that are uttered by a user are converted into a format that can be read by a computer via speech recognition software. This allows the user to simply operate the device by speech. Automatic voice recognition, also known simply as speech recognition, is another name for speech recognition (ASR),for individuals diagnosed with autism spectrum disorder

(ASD), language impairment (LI), or both. The goal of the research is to improve the accuracy of the system.

There is a significant amount of ambiguity surrounding the concept of a "normal voice." Because there is a continuum between a normal voice and a disordered voice, it is difficult to define the basic characteristics of this vocal condition. A normal voice is one that has a quality that is generally ordinary and that enables acceptable communication without requiring an excessive amount of effort or pain. A voice that has an abnormal quality [5] , known as hoarseness, may be rough, harsh, breathy, weak, or strained. Hoarseness is a phrase that represents this abnormal quality. According to the World Health Organization (WHO) [5], a voice problem, also known as dysphonia, is any impairment, limitation in activity, or restriction in participation that arises as a direct consequence of a structural or functional abnormality of the voice system[6]. Purposes For the purpose of testing the incremental validity of more expensive vocal development variables relative to less expensive variables for predicting later expressive language in children with autism spectrum disorder, this study was designed (ASD). We pay special attention to the additional value of coding the quality of vocalizations rather than the number of vocalizations since coding quality adds expenditure to the process of coding[6]. Children who have autism spectrum problems (also known as ASC) have significant challenges in reading and reacting appropriately to the emotional and mental states that can be deduced from the facial expressions of others. These issues in empathy are the root cause of their difficulties in social communication, which are the foundation of the diagnostic. In this article, we explore the question of whether or not young children with ASC may be taught components of empathy.

The display of emotion is a key component of all forms of communication as well as a component that is essentially present in all aspects of human behavior. Children who have autism may have unique causes for their emotional outbursts and may convey their thoughts in unusual ways, but autistic children nonetheless feel the same things that everyone else does. These children may articulate their thoughts in unusual ways. On the other hand [7], younger people frequently struggle to articulate their emotions and may need guidance in order to do so effectively. As a result of a wide variety of stimuli, children are capable of feeling a wide range of emotions; the objective of this activity is to develop the ability to recognize these emotions in children by listening to what they have to say. Because of this, it is possible to predict the appropriate emotion based on children's attitudes and the ease with which children are able to communicate their feelings to others. In this context, children express a wide range of feelings in response to various circumstances, including happiness, sadness, anger, and pain [7].

Therefore, this study is about the implementation of an Autism children application to Enhance their state and can be implemented using multiple machine learning algorithms and classification techniques to provide the highest accuracy and best solutions for users to recognise them.

II. BACKGROUND

In past years, several systems have been proposed for Autism children Recognition. Different machines for Autism children' Recognition have been used. Below are some articles we have reviewed on the Autism children' voice and emotion recognition application.

This study provides[1], an overview of the most recent developments in neural speaker diarization techniques. In addition, the papers Discuss the ways in which speaker diarization systems have been merged with voice recognition applications, as well as the ways in which the recent boom in deep learning is pointing the way towards to the combination of these two components so that they are similar to one another. In particular, the research focuses on how deep learning is paving the way toward simultaneously modeling speaker identification applications and systems for voice recognition. It is intended to be a beneficial contribution to the community by offering a survey work by merging the latest breakthroughs using neural approaches, and by doing so, it will facilitate further progress toward a speaker diarization system that is more effective. The majority of the approaches that are discussed in Section 2 [1] of the proposed taxonomy are those that fall under the category of "Non-diarization objective." These techniques are employed in conventional, modular speaker diarization systems.

Within the scope of this research [3,4], the development and enhancement of an automatic speech recognition (ASR) system for children diagnosed with autism spectrum disorder and ranging in age from 6 to 9 years is investigated (ASD). Working with only 1.5 hours of target data in which children perform the Clinical Evaluation of Language Fundamentals Recalling Sentences task, the authors of this paper used deep neural network (DNN) weight transfer techniques to adapt a large DNN model that was trained on the LibriSpeech corpus of adult speech. The target data was collected while the children were working on the Clinical Evaluation of Language Fundamentals Recalling Sentences task. The primary objective is to locate the optimal proportional training rates for each layer of the DNN. A word error rate of 29.38% is achieved with the optimal arrangement (WER). The researchers augment the training with portions of the OGI Kids' Corpus, adding 4.6 hours of typically developing speakers aged kindergarten through third grade. Using this configuration, the researchers explore the effects of quantity and similarity of data augmentation in transfer learning. In addition, they augment the training.

In order to investigate the incremental validity of more expensive vocal development variables in comparison to less expensive variables for predicting later expressive language in children with autism spectrum disorder, this study was designed [6]. (ASD). Because coding quality adds additional price to the process of coding, researchers focus a lot of their attention on the additional value that comes from coding the quality of vocalizations rather than the quantity of vocalizations. They are also interested in the added value of human-coded voice variables, which are more expensive than those derived by computer analyses. Method Participating in the study were 87 children diagnosed with ASD who were between the ages of 13 and 30 months. One variable was produced via human coding of brief communication samples, and the other was derived from an automated procedure for daylong naturalistic audio samples. Both variables were used to determine the quantity of vocalizations.

In this scientific study [7,8], the authors investigate whether or not young children with ASC are capable of learning parts of empathy. They discuss a research that analyzed The Transporters, an animated series aimed to improve children's understanding of emotions who have autism spectrum disorder (ASD). Children diagnosed with ASC ranging in age from 4 to 7 years old watched The Transporters each day for a total of 21 days. The participants were evaluated on their emotional vocabulary as well as their ability to recognize a range of emotions at three different degrees of generalization before and after the intervention. On every level of the task, the performance of the intervention group was comparable to that of typical controls at all times. This improvement was much greater than that seen in a clinical control group.

By considering these factors we recommend a Mobile-application as a solution and hope it will be of help to Users who are not able to identify Autism children' differences and who do not have the facilities to obtain the best and free solutions to this problem, can access our mobile application and get the help they need.

III. METHODOLOGY

Fig.1 shows a block diagram of the Integrated Automatic Speech Recognition System to Analyze Autism spectrum disorder in young children: A Supporting application for Enhancing the Analyze Autism spectrum disorder. The application's primary goal is to enable support for Analyze Autism spectrum disorder as a helping tool.

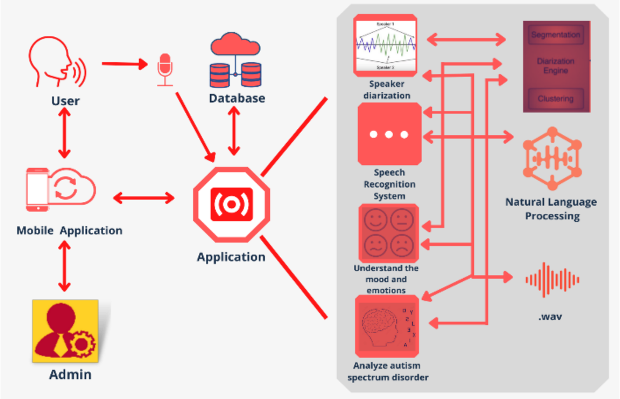
User
Database
Mobile Application
Application
Admin
Speaker diarization
Segmentation
Diarization Engine
Clustering
Speech Recognition System
Natural Language Processing
Understand the mood and emotions
.wav
Analyze autism spectrum disorder

Figure 1: System Architecture

The Proposed system will be carried out under four main components.

1) Speaker diarization Using Language based.
2) Speech Recognisation System.
3) Understand the mood and emotion.
4) Analyse autism spectrum disorder
5) Give the solution to Analyze the Autism spectrum using a mobile app.

Different types of CNN algorithms and models will be used to examine the data. After the ML (Machine Learning) model has been trained, the binary classifier results will be categorized. Python was chosen as the programming language, along with the libraries, Deep Learning, ML framework, K-means clustering, Random forest classifier technique Model to Implementation, and Feature extraction for Analyze the Autism spectrum.

1) Speaker diarization Using Language based

This paper represents the implementation of the study topic associated with automatic voice diarization, which is a field that requires speaker diarization as an upstream processing step. However, in recent years, speaker diarization has developed into an important key technology that may be utilized for a variety of activities, including navigation, retrieval, or higher-level inferences based on audio data. As a consequence of this, numerous significant advancements in terms of accuracy and robustness have been documented in journals and conferences pertaining to this field. Some of these challenges include overlapping speech, having access to many microphones, and dealing with multimodal information. Finally, this paper presents an analysis of speaker diarization performance among adults and children, Using a Random Forest Classifier and CNN techniques.

Secure and user-friendly application system based on an audio processing system that will allow users to Analyze Autism spectrum disorder using their mobile. Speaker Diarization is the technical method of dividing a single audio recording stream into several identical speakers. These categories pertain to every single speaker.

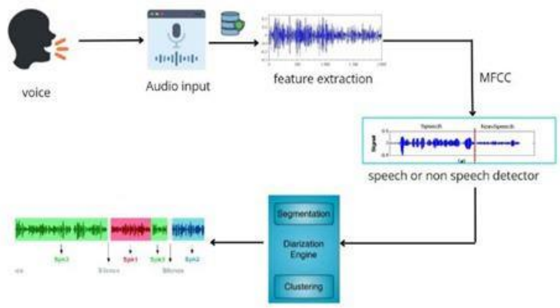

Figure 2: Speaker diarization Architecture

Above Fig. 2 is the individual system diagram where this component is represented as a small box, and the middlebox in the research component is typical for everyone; hence it involves machine learning. This involves input:

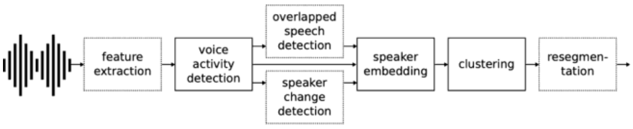

Figure 3: A typical diarization pipeline

Speaker Detection: In this step, intelligence is used to distinguish speech from audio recordings and sound effects.

This phase entails extracting tiny segments from the audio file. Typically, each speaker has a segment that lasts around one second.

Embedding Extraction: This stage integrates and produces the neural network for all embedded speech segments that have been constructed and gathered. These embeds can be translated into different data types, as well as text, photos, and documents. These distinct data kinds can be utilized inside a framework for in-depth learning.

Clustering: After embedding the sections, as seen in step three, that this next step is to cluster the bundles. After clusters have been created, they are often labeled according to the number of speakers present..

Transcription: Finally, we reach the transcription step. Once the clusters are created and properly labeled, the audio can be split into separate clips for each speaker. Those clips will be sent through a speech-to-text application or speech recognition system that will eject the transcription.

a. Speaker Diarization Pipeline

A speaker diarization system consists of a Speech Signals Detection (VAD) model to obtain the timestamps of audio where speech is being said while ignoring ambient noise and a Speaker Embedding extractor model to obtain speaker embeddings on speech segments obtained from VAD time stamps. These models are intended to extract the timestamps of spoken audio while disregarding background noise. Then, these speaker embeddings would be grouped in accordance with the total number of voices in the audio recording.

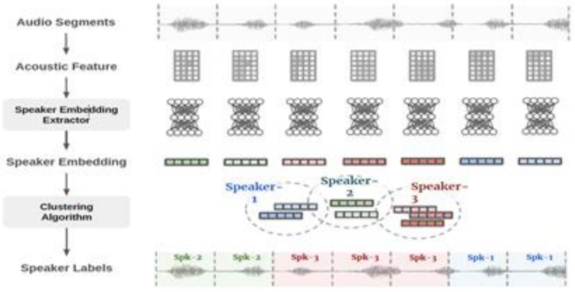

Figure 4: Speaker Diarization Pipeline

b. MFCC

The feature is the name of the acoustic property that the speech signal possesses. The technique of extracting a little quantity of data from the speech signal so that it can be utilized in the future to represent each individual speaker is referred to as feature extraction. There are a variety of feature extraction methods available, however the Mel Frequency Cepstral Coefficient (MFCC) is the method that is most frequently utilized. In this work, the data retrieved from the speaker speech signal is used to recognize the emotions being expressed by the speaker. The Mel Frequency Cepstral Coefficient (MFCC) method is utilized in order to determine an individual's emotional state based just on their voice. Children on the autism spectrum were used to validate the devised system, and the results indicated that it was approximately 80% effective.

2) Speech Recognition System Using NLP.

The suggested method will use audio processing in a neural network and a trained deep-learning model for Speech Recognition System. In this section, we concentrated on the functions of interpreting the language content and obtaining samples of people's speech. The pre-processing and feature extraction step come next in the process. The method of MFCC feature extraction is the one that we will use in this situation. The next stage of the procedure involves selecting the word with the highest likelihood using HMM technology. The final part of the process involves parsing and attributing the results. And that is it. The process is something that comes after the techniques. As a result, we are able to comprehend the meaning of the language.

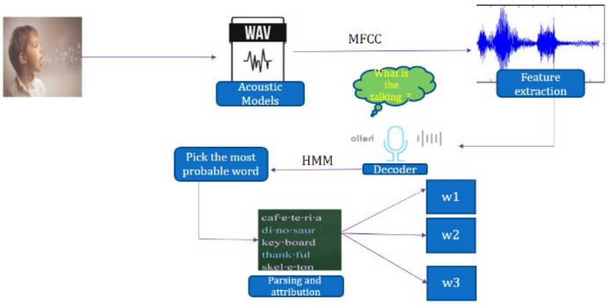

Figure 5: Speech Recognition System

The PyAudio module, which is required for speech recognition, is made available by Speech Recognition, which also makes the process of providing audio input very simple. With SpeechRecognition, you won't have to spend hours crafting scripts from scratch in order to gain access to microphones and process audio files; instead, you'll be up and running in a matter of minutes. The flexibility of the SpeechRecognition library can be attributed to the fact that it serves as a wrapper for a number of well-known speech APIs. The Google Web Speech API is one of these, and it is the only one that allows a default API key that is hard-coded into the SpeechRecognition library.

Speech recognition, also known as automatic speech recognition (ASR), computer voice recognition, and speech-to-text, is a technique that turns human speech into text. Voice recognition is a common application of the technology; unlike speech recognition, which focuses on converting verbal to written speech, voice recognition seeks to recognize the distinctive voice of each individual user. These models work exceptionally well with problematic machine learning models to provide apparent gains. Hidden Markov models, which are regarded as the standard approach to automated speech recognition (ASR), have been enhanced by advancements in ASR over the course of several decades.

a. NLP

In this paper, we use NLP in terms of the development of speech recognition systems, natural language processing (NLP) is far more important than guided dialogue when evaluating the growth of ASR technology. A typical NLP ASR system will include a database including at least sixty thousand words. With the addition of a three-word sequence, the total number of possible word combinations approaches 215 trillion! The algorithm is designed to mimic, to some extent, how people interpret speech and behave accordingly.

Going into more detail and taking one step at a time, we believe that natural language processing (NLP) serves primarily as a means for a very important aspect that is referred to as "Speech Recognition." In this process, the systems analyze the data in the form of words, whether they are written or spoken. Our assistance with the systems that convert speech to text and text to speech. In our opinion, we will center our attention on speech-to-text technology because it will enable us to use audio as a major source of data and then train our model using deep learning.

b. Random Forest classifier

This speaker recognition system uses a Random Forest as a classifier to distinguish between all of the different speakers. Speakers employ the MFCC and RPS approach for the purpose of feature extraction. The results were generated by using the Random Forest. The results of the classifier look promising. It has been noticed that the level of accuracy in MFCC is much higher than compared to if we compare this method to the RPS approach.

3) Understand the mood and emotion.

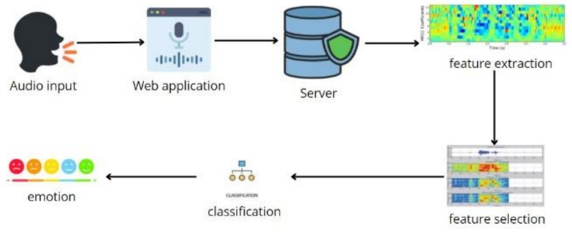

Figure 6: Understand the mood and emotion

Autism-related emotions may have specific triggers, yet autistic children feel the same as everyone else. Children may need help expressing their feelings. Listening to children's voices can help us recognize their emotions. In this situation, children exhibit a wide range of feelings, including happiness, sadness, anger, and pain. It is feasible to forecast the appropriate emotion based on children's attitudes and the ease with which they can transmit their sentiments to others.

a. Deep Convolution Neural Network Design

Convolutional and pooling layers are typically combined for computation in the CNN network structure. After the convolutional layer extracts the features, the pooling layer performs the dimensionality reduction operation by selecting the standout features from the amplified feature attributes before passing the data to the fully-connected layer for classification. The authors of [8] proposed a novel CNN architecture with special strides rather than a pooling scheme to extract the important high-level features from spectrograms of speech signals for the purpose of down-sampling the feature maps rather than the pooling layers. This was done in order to complete the task described in [8]. Data preprocessing was undertaken in the study that was referenced in [8], and this involved the authors removing noise from the data using a unique adaptive thresholding technique, followed by the elimination of silence sections of aural data. Using a method known as five-fold cross-validation, the authors of the study conducted utterance-based experiments on SER. The data were divided so that 80% were used for training the model and 20% were used for testing the model. The data were split using the 80/20 rule. In addition to that, there was no data enhancement done.

b. Method for Detecting Keywords

Because it involves selecting words from inside the text, this technique is both easy to put into practice and obvious in its operation. The difficulty of locating keywords from a given data collection as substrings in a given string is referred to as the keyword pattern matching problem. These words can be categorized as repugnance, melancholy, joy, rage, terror, surprise, and many more emotions. After the sentences have been tokenized, the individual words are then matched with the various types of emotions by taking into consideration the strength and negation of the words. It is possible to identify the type of emotion, as well as evaluate the individual's actions.

c. Method of Lexical Affinity

The probabilities that are determined by using this approach are incorporated into language corpora. This has some drawbacks because the probabilities that are assigned are biased toward corpus-specific metaphors that are found in texts. They are unable to recognize the feelings that are conveyed in the text since those feelings do not exist on the word level where this method functions.

d. Learning-Based Approach Learning-based methods

Attempt to distinguish feelings in view of recently prepared results and classifiers. These outcomes and classifiers are planned with an assortment of AI classifiers, for example, support vector machines, explicit factual learning strategies, and choice trees, to figure out which feeling class the text has a place with. As a result of limited feature collecting, our approach faces the challenge of only being able to classify phrases into two categories: positive and negative.

e. An Approach That Makes Use of Both

Combining the method that is based on keywords with the way that is based on learning results in an approach that provides correct results and successfully handles the high costs associated with information retrieval jobs. Despite the fact that it is highly effective, these lacks concealed emotional patterns as well as an in-depth study of context and semantic components.

4) Analyse autism spectrum disorder

The purpose of this study is to determine, based on a child's linguistic or communicative skills, whether or not they fall into the "normal" or "abnormal" category. It is crucial to explain one's requirements and develop a relationship using speech because it is one of the most fundamental kinds of communication. People need to have a holistic comprehension of the topic in order for them to be able to comprehend language. They are also required to form an image in their heads of the message to which they are responding. The primary concentration is on the speech recognition feature because one of the most important objectives is to comprehend the language being used. As a consequence of this, in this case, extracting sentences from recordings of children's voices after such recordings have been made. Therefore, our system is able to readily identify the phonemes through children's voice clips, and ultimately, we are able to see the phonemes on a particular monitor, which enables us to easily recognize the linguistic content through children's voice clips. as well as those on the autism spectrum.

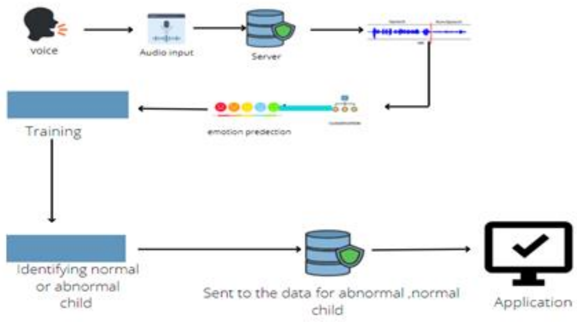
voice
Audio input
Server
Training
emotion predection
Identifying normal or abnormal child
Sent to the data for abnormal ,normal child
Application

Figure 7: Analyse autism spectrum disorder

After median smoothing filtering, short-term energy and spectral centroid curves are smoothed. In double-threshold, the threshold is dependent on experience. However, different people or situations have distinct speech qualities, thus utilizing the same threshold yields inaccurate results.

A new approach improves noise detection accuracy. This algorithm dynamically determines the threshold. First, compute the smoothed feature sequence's histogram. Histograms precisely describe data distribution and estimate variable probability distribution. First, partition the range of values at equal intervals and count the data in each piece. This will build the histogram. Start by finding the minimum and maximum spectral centroid characteristic coefficients. Next, the minimum-to-maximum range is averaged into L sections. Before generating the histogram, one counts the spectral centroid coefficients in each region. Assume the histogram item's value. The local maximum value M of the statistical histogram is attributable to the fact that if the likelihood of the occurrence of the character sequence is much higher than that of the neighboring location, then it is likely that the position is the transition from nonspeech to speech. This explains the statistical histogram's local maximum M. If a segment appears in the histogram more than its neighboring segments, the middle of the segment's characteristic coefficient is a local maximum.

IV. RESULTS

Researchers have access to a large database known as the Autism Kids' Speech Corpus. This data was used to train a speech recognizer for children's speech that is independent of the speaker as well as the vocabulary. In this section of the study, we describe the recognition accuracy achieved by using CNN, and MFCC classifiers in our analysis. Every classification result is arrived at after ten separate rounds of cross-validation. Simple MFCCs were used as the building blocks for the neural network that was used in this experiment. The importance of the speaker normalization (SN) step, which comes before recognition, is investigated. SN is useful because it can adjust for differences in speech that are attributable to factors other than shifts in an emotional state. The value of the threshold is determined by using the local maximum of the histogram of the statistical feature sequence.

V. CONCLUSION

The primary focus of this study is on the voice detection of Autism children's identification systems. For example, we are the using MFCC feature extraction model, and then we have to use the csv training model. Therefore, we have to store the training models and convolution neural network (CNN)technology for identifying and recognizing objects from speech and predicting the correct object. There are many tools we are using, android studio, flutter, python, and PyCharm. And finally, our main concept was machine learning. We are using this concept so that it can more accurately predict the appropriate outcomes without being explicitly programmed.

REFERENCES

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe & S. Narayanan. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language, 72*, 101317. DOI: 10.1016/j.csl.2021.101317.

[2] A. Gorodetski, I. Dinstein & Y. Zigel. (2019). Speaker diarization during noisy clinical diagnoses of autism. *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2593-2596. DOI: 10.1109/EMBC.2019.8857247.

[3] A. S. Shinde & V. V. Patil. (2021). Speech emotion recognition system: A review. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3869462.

[4] R. Gale, L. Chen, J. Dolata, J. van Santen, & M. Asgari. (2019). Improving ASR systems for children with autism and language impairment using domain-focused dnn transfer techniques. *Interspeech*. DOI: 10.21437/interspeech.2019-3161.

[5] P. J. Bradley. (2010). Voice disorders: classification. *Otorhinolaryngology, Head and Neck Surgery*, pp. 555–562. DOI: 10.1007/978-3-540-68940-9_60.

[6] J. McDaniel, P. Yoder, A. Estes & S. J. Rogers. 92020). Predicting expressive language from early vocalizations in young children with autism spectrum disorder: which vocal measure is best?. *Journal of Speech, Language, and Hearing Research, 63*(5), 1509–1520. DOI: 10.1044/2020_jslhr-19-00281.

[7] S. Baron-Cohen, O. Golan & E. Ashwin. Can emotion recognition be taught to children with autism spectrum conditions?. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), pp. 3567–3574. DOI: 10.1098/rstb.2009.0191.

[8] L. Trinh Van, T. Dao Thi Le, T. Le Xuan & E. Castelli. (2022). Emotional speech recognition using deep neural networks. *Sensors, 22*(4), 1414. DOI: 10.3390/s22041414.