MACHINE LEARNING BASED AUTOMATED SPEECH DIALOG ANALYSIS OF AUTISTIC CHILDREN

Anjana Wijesinghe
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
anjana.w@sliit.lk

Pradeepa Samarasinghe
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
pradeepa.s@sliit.lk

Sudarshi Seneviratne
Faculty of Medicine
University of Colombo
Colombo, Sri Lanka
sudasene@yahoo.co.uk

Pratheepan Yogarajah
School of Computing, Eng & Intel. Sys
Ulster University
Ulster, United Kingdom
p.yogarajah@ulster.ac.uk

Koliya Pulasinghe
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
koliya.p@sliit.lk

***Abstract*—Children with autism spectrum disorder (ASD) have altered behaviors in communication, social interaction, and activity, out of which communication has been the most prominent disorder among many. Despite the recent technological advances, limited attention has been given to screening and diagnosing ASD by identifying the speech deficiencies (SD) of autistic children at early stages. This research focuses on bridging the gap in ASD screening by developing an automated system to distinguish autistic traits through speech analysis. Data was collected from 40 participants for the initial analysis and recordings were obtained from 17 participants. We considered a three-stage processing system; first stage utilizes thresholding for silence detection and Vocal Activity Detection for vocal isolation, second stage adopts machine learning technique neural network with frequency domain representations in developing a reliant utterance classifier for the isolated vocals and stage three also adopts machine learning technique neural network in recognizing autistic traits in speech patterns of the classified utterances. The results are promising in identifying SD of autistic children with the utterance classifier having 78% accuracy and pattern recognition 72% accuracy.**

***Index Terms*—Autism Screening, Speech Deficiencies, Neural Networks, Audio Classification, Speech Pattern Recognition**

I. Background

ASD forms a spectrum of disorders characteristing social and communication difficulties, stereotypic and repetitive behaviours [1], [2]. The Centers for Disease Control and Prevention of Autism and Developmental Disabilities Monitoring Network (ADDM) report that approximately one in 68 children can be identified with the above symptoms [3].

South Asia region represents more than 20% of the worlds population, yet the prevalence of ASD in this part of the world is still largely unknown [4]. In Sri Lanka, the only community based study done reported 374 children aged 18-24 months who were initially screened for autism, using Red Flag criteria, and then diagnosed using DSM-IV criteria, resulted in 4 (1.07%) diagnosed as ASD [5]. According to the review done by Mohammad Didar Hossain et al, Sri Lanka reported the highest prevalence from the South Asian Countries but consistent with the rates in the Western world [4].

In a study done in Sri Lanka it was revealed that by the age of 24 months, only 14.3% of children with autism have sought treatment [6], which was considerably low compared to Western countries. One of the most significant challenges faced by individuals with ASD and their families is difficulty in obtaining a diagnosis of ASD. Currently, there are no autism screening tools that are widely used in Sri Lanka among pediatricians that would assist with the identification process [5]. A study done in a tertiary children hospital revealed that 34% of the doctors were unaware of the main presenting symptoms as speech delay and a further 39% failed to recognize the comorbidities in ASD [7].

Many studies have suggested that autistic children suffer speech delays or language deficiencies either through empirical or anecdotal evidence [8], [9]. In Sri Lanka also, [6] showed that speech and language delay was the commonest concern (82.4%) expressed by the parents as the presenting symptom. Specifically, [9] shows children identified at risk for ASD shows high sensitivity in communication delays. Most prominent of these are the delay in first words or phrases and expressing themselves in a sentence or a grammatical utterance [10]–[12]. Further the vocabulary of children suffering from ASD are far more simple and limited when compared with a typically developing child [12], [13].

Our research relies on machine learning to identify autistic

The research was funded by the research grant (grant no. FGSR/RG/2017/05) provided by Faculty of Graduate Studies and Research, Sri Lanka Institute of Information Technology, Sri Lanka.

traits in developing children. Though there are few research studies [14] on automated vocal analysis, they cannot be directly applied to Sri Lanka due to significant differences in language [15], culture and facilities recommending the need of a special study on this area. The research directly contributes to increasing the autism diagnosis rate at early stages, while providing a cost efficient system to reach the remote areas. A rich database consisting of autistic symptoms analysis data at early stages and audio recordings of autistic and typically developing children will be created as collateral.

The research paper is organized as follows: section II details the research and analysis carried out, section III describes the data collection process, section IV details the implementations carried out, section V shows the results of the research while section VI discusses the outcomes and the future direction.

II. Autistic symptoms analysis

As the first research step, surveys were carried out. Data related to identifying the first autistic symptom was collected from parents of children diagnosed with ASD.

As shown in Table I, data collected from 40 participants were categorized based on the first symptom noticed by the parent. Majority of the parents have identified ASD through speech deficiency (SD) symptoms such as speech delay, speech regression, communication problems and echolalia. The rest were identified on their poor social interaction (PSI) such as poor eye contact, unresponsive to name, repetitive behaviours and isolation.

TABLE I
Overview of data collected.

	PSI	**SD**
Percentage	27.59%	72.41%
Gender	F:12.50% M:87.50%	F:23.81% M:76.19%
Noticed age	2.05 years	2.34 years
Diagnosed age	2.88 years	2.60 years
Treatment age	2.90 years	2.65 years
Therapy time	1.29 hours	1.74 hours

The average age ASD was identified with SD is higher than PSI, which might be due to parents being reluctant to consult a doctor as delay in speech is common to even about 20% of typically developing children. The gap between diagnosed and treatment started age shows that once diagnosed, the children were treated immediately which should be appreciated with existing limited medical facilities. Therapy time in Table I shows children spend more than an hour for treatments per month at medical facilities. As medical facilities are unreachable or requires travelling long distances for many, an easily accessible mobile ASD screening and intervention facility has become essential.

As SD is the major symptom and noticed at later ages compared to PSI, our research was focused on identifying SD symptoms detected at early stages of autism.

A. Speech deficiencies

Focusing the research identified a small set of main SD categories as blabbering, neologism and echolalia. Table II shows the extension of the survey to speech based analysis where blabbering and neologism are observed to be common but echolalia not as prominent.

TABLE II
Speech based analysis.

Blabbering		83%
Neologism		72%
Echolalia		55%
Vocabulary size		
None 29%	Limited 61%	Mediorce 10%
Sentences		38%
Conversation		3%
Unnatural pitch and tone		31%
Response to name		
Yes 17%	Varying 45%	No 38%

Children able to speak with varying vocabulary accounted for the least and a majority of the population weren't able to speak in sentences with more than 3 meaningful words while typical autistic children weren't capable of holding a simple conversation. Few children were observed to use a non-native pitch and tone, which is reasonably common among autistic children [16], [17]. A majority of the population only responds to their name depending on the situation (when they want something or called a few times).

Interestingly it was noted that all the children interviewed had one or more SD symptom regardless of first noticed symptom being PSI. These results emphasis the need for a detailed research and a system to screen SD. As the first step towards that objective, audio recordings were collected.

III. Audio data collection

In collaboration with Lady Ridgeway Hospital (LRH) for children in Colombo, Sri Lanka, and by securing the ethical clearance from the facility to obtain recordings; audio recordings were obtained. The LRH for children is a tertiary care children hospital and is considered the largest children hospital in the world. A voice recorder was placed either within a pocket in the child's clothing or within a meter from the child for periods varying from 2 to 10 hours. Identical voice recorders were used for all the data collection. The recordings were of conversations with a familiar adult in a familiar environment for the child. These conditions were necessary to make the data more real and unbiased.

Data was gathered from autistic and typically developing children of ages between 1.5 and 3 years with equal number of participants to create an unbiased dataset, 8 autistic children and 9 typically developing children. The audio data obtained is analyzed to identify autistic traits in speech.

IV. Audio analyzing system

Audio recordings are segmented, categorized, labelled and analyzed to identify speech patterns in the autistic and control samples to develop the screening tool for SD.

A. Audio segmentation and silence filtering

The audio cannot be directly analyzed as it could contain mutually exclusive information and due to the varying duration. We consider the silences and maximum allowed audio length to segment the audio to a fixed length. Energy level of the audio varies with the amount of sound present; thresholding is applied to the energy level of the audio parting low energy clusters (LEC) as silence and high energy clusters (HEC) as non-silent, represented by Figure 1.

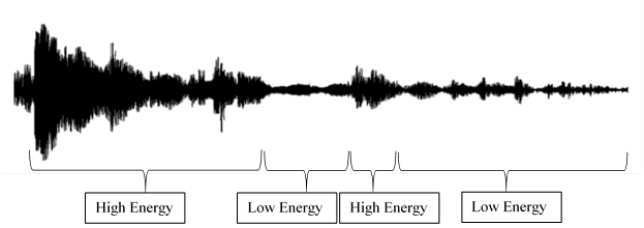

Fig. 1. Identifying HEC and LEC.

Even though the average duration required for an utterance of one or two syllables is considered 600ms [14]; analyzing utterances of autistic children blabbering or cooing shows energy level of the audio remains high for longer than 600ms. Increasing maximum segment length increases chance of interruptions (unwanted noise) being present and containing utterances from more than one person. Empirically, 2s was selected as optimal maximum segment length granting the maximum exposure to utterances and minimum presence of noise and interruptions.

A recurrent algorithm is used to segment long audios. During the first iteration, the beginning and end of LEC longer than 2s are marked by thresholding. The audio is segmented at the marked points labelling the LEC as “Silent” and HEC shorter than 2s are labelled as “Non-silent”, as shown by 2s iteration in Figure 2. Unlabelled segments are passed onto the next iteration where LEC longer than 1s are marked using thresholding, then segmented labelling LEC as “Silent” and HEC shorter than 2s as ”Non-silent”. The process is repeated for 500ms, 200ms, 100ms and 50ms as shown by Figure 2. HEC longer than 2s present after the completion of the 50ms iteration are discarded. Vocals filtration is then applied to the segments labelled “Non-silent”.

B. Vocal filtering

Human vocal frequencies are constricted to a specific range, generally 80Hz to 4000Hz are considered vocal frequencies [18]. Thereby audio segments containing no vocal frequencies or very few vocal frequencies for a very short duration could be considered noise only segments.

Using voice activity detection (VAD) [18], duration of vocal frequencies occurring is identified. The ratio of the vocal to non-vocal frequencies duration (V-NV) is calculated per segment using the identified duration and total duration. Thresholding is applied to classify the segment as vocal or non-vocal which was empirically set to an optimum of 0.5, allowing a maximum of actual vocal to be identified precisely.

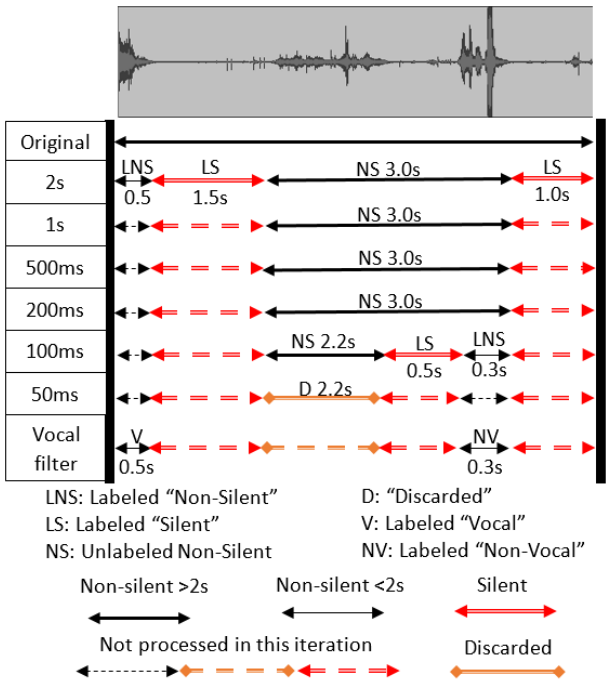

Fig. 2. Audio segmentation process

The vocal filter relabels vocal audios to “Vocal” and non-vocal to “Noise”, represented by the vocal filter iteration in Figure 2. Vocal audio segments are classified further to identify speech patterns.

C. Utterance classification

Basic speech patterns identification requirements were narrowed down to the 7 categories; child uttering a meaningful word, child uttering a meaningless word, vegetative sounds made by the child, adult utterances, noises, silences, auxiliary (belonging to more than one category). Initially, human transcribers relabelled the vocal audio segments to the relevant categories to be used as training data.

Machine learning technique neural networks (NN) are excellent at classifying data by identifying patterns. Convolutional NN architecture is widely used for image classification for its effectiveness [19]. This architecture was selected to classify the audio segments into categories [19], audio is transformed to mel-frequency cepstral coefficients (MFCCs) giving attributes similar to a data representation of an image which is analyzed and classified. MFCC transformed audio are perfect input to convolutional NN [19]; traditional Fast Fourier Transformation (FFT) has a linear resolution and only determines the frequency content whereas MFCCs contain features to how humans perceive pitch [20].

A 7 category utterance classifying model was developed to label audio segments (NN-1) consisting of 5 layers. The inputs are 58 frames of 20 MFCC features generating 1160 units which are convoluted to the first hidden layer of 256

rectified linear units (ReLU) by applying a 2*2 convolution window. The rest of 3 hidden layers have 128, 64 and 32 ReLU respectively in conjunction with dropouts providing the optimal combination for hidden layers [19]. Output layer have 7 softmax units corresponding to the categories. Softmax presents probabilities class-wise thus is particularly suitable for multi-class scenarios, true class is elected by the highest probability.

After NN-1 classifies and labels the audio, length of the audio segment is appended to the label creating a dataset per child to be analyzed to determine SDs present.

D. Speech pattern classification

Speech pattern classifying model was developed to distinguish autistic from typically developing children's speech patterns (NN-2). The created dataset is analyzed for patterns, including the speech rate of the child with contrast to the adult, the total silence within the conversation and the total duration the child speaks. The initially conducted analysis proves autistic children lack in these parameters compared to typically developing children, NN-2 employs these to distinguish between autistic and typically developing. The duration of each utterance category (except auxiliary), total child duration and total audio duration per 10 minutes are fed into NN-2 with the label (autistic or typical) as training data. Total audio duration is used as a control input as the total audio duration is not exactly 10 minutes. A optimal input time duration of approximately 10 minutes was deduced empirically as NNs require fixed input lengths giving consideration to time required for a conversation and limited data available. Longer datasets are sliced to create multiple smaller dataset, which indirectly increases the performance of NN-2 as the amount of training data is drastically increased.

NN-2 consists of 6 layers. Inputs are approximately 10 minute frames of 8 features: 6 categories (auxiliary category is neglected), total child duration and total audio length, generating 8 units. 5 hidden layers have 128, 64, 32, 16, 8 ReLU with dropouts respectively. The output layer has one sigmoid unit producing the probability the speech pattern belongs to either autistic or typical.

V. Results & Discussion

The segmentation algorithm successfully segments 80% of the recording duration to segments less than 2s duration, the rest are discarded. The filters are capable of labelling approximately 90% and 30% of silent and noise segments respectively. Unfamiliar data was used as test data for both NNs in order to obtain an unbiased output.

NN-1 achieved a training accuracy of 79% and testing accuracy of 78%. Individually, silence class had the best relationship between precision and recall followed by the adult and noise, whereas vegetative and meaningful classes had inferior relationships. Amount of training data per class was unbalanced; comparatively, noise class accounted for the highest followed by silence and adult whereas the vegetative and meaningful contained data equal to approximately 20% of amount of data in the noise class. The amount of training data is directly related to the precision of each class, contributing to the average precision likewise. Overall, Figure 3 shows a good relationship between precision and recall on average, thus NN-1 is effective at classifying audio segments with a high precision of 86%. Similar research done for English language [14] achieves a similar accuracy for the utterance classifier.

NN-2 achieved a training accuracy of 90% and test accuracy of 72%. An average precision of 58% achieved is low and the relationship between precision and recall throughout is observed to be inadequate as shown in Figure 4. Due to the limited training data available, even with fairly high accuracy, the low precision makes NN-2 fairly ineffective.

VI. Conclusion

An automated autism screening tool would assist in increasing the diagnosis rate while no similar research has been conducted in Sri Lanka, proves the necessity for this research. We surveyed the ASD symptoms noticed at early ages, identifying SD as the most prominent. Recordings of autistic and typically developing children were obtained and analyzed for speech patterns incorporating machine learning and audio processing techniques. The recordings were captured at the child's natural environment, allowing the child to behave naturally while also capturing the local culture traditions and language patterns.

Overall, the results were promising. A highly effective utterance classifier was implemented, but speech pattern recognizing was fairly ineffective. In comparison, NN-1 was trained with significantly more data than NN-2, proving NN-2's performance could be further enriched with more data. The unbalance of the training data should be addressed. Due to the nature of the data, significantly increasing sample count in the low data classes requires the total data size to be increased by a minimum of 5 times. The limited data set is mainly due to cultural beliefs and traditions in Sri Lanka, as a majority of the parents are not open in sharing the data.

Next stage of this research is to gather more data around the country creating a more unbiased dataset in-turn developing a more effective NN. In addition, a mobile application will be developed to screen for autism related SD even by parents, which is proposed to be extended to include intervention.

References

[1] L. Wing and J. Gould, "Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification," *Journal of autism and developmental disorders*, vol. 9, no. 1, pp. 11–29, 1979.

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed., 1994.

[3] Centers for Disease Control and Prevention (U.S.), "Prevalence of autism spectrum disorder among children aged 8 years autism and developmental disabilities monitoring network, 11 sites, united states, 2010," *Morbidity and Mortality weekly report: MMWR*, vol. 63, no. 2, pp. 1–21, 2014, http://www.cdc.gov/mmwr/pdf/ss/ss6302.pdf.

[4] M. D. Hossain, H. U. Ahmed, M. M. Jalal Uddin, W. A. Chowdhury, M. S. Iqbal, R. I. Kabir, and M. Sarker, "Autism spectrum disorders (ASD) in south asia: a systematic review," *BMC Psychiatry*, vol. 17, p. 281, 2017.

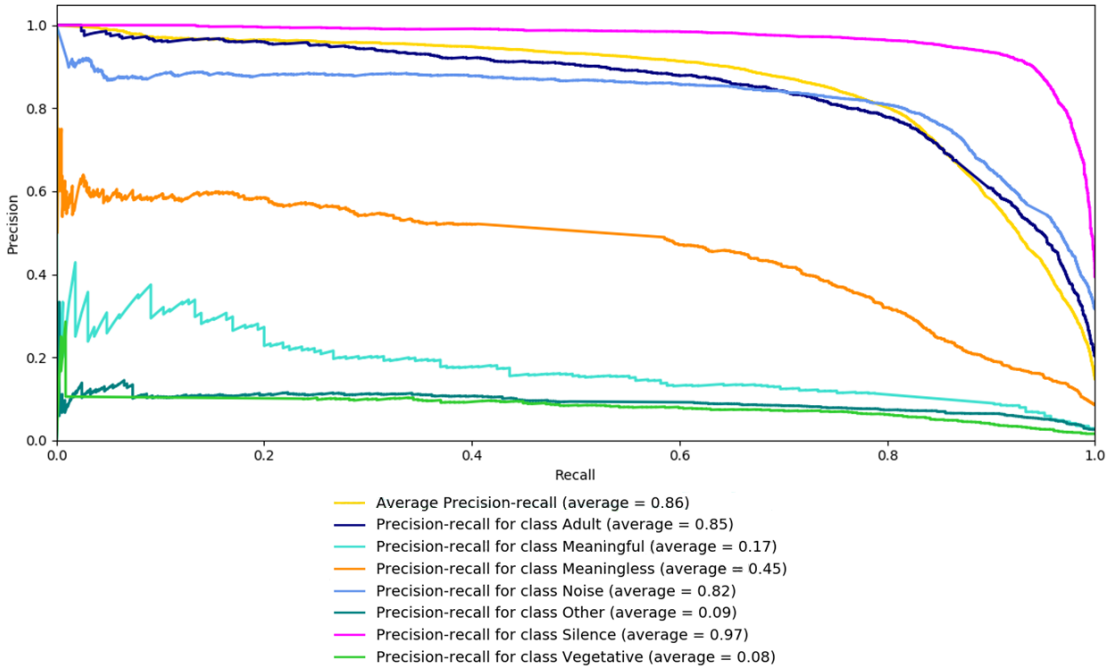

Fig. 3. Precision Recall curve for test data of NN-1.

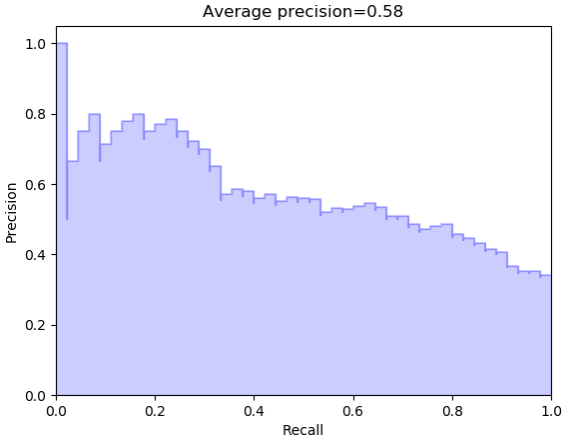

Fig. 4. Precision Recall curve for test data of NN-2.

[5] H. Perera, K. Wijewardena, and R. Aluthwelage, "Screening of 18-24-month-old children for autism in a semi-urban community in Sri Lanka." *Journal of tropical pediatrics*, vol. 55, no. 6, pp. 402–5, 2009.

[6] H. Perera, K. Jeewandara, C. Guruge, and S. Seneviratne, "Presenting symptoms of autism in Sri Lanka: analysis of a clinical cohort," *Sri Lanka Journal of Child Health*, vol. 42, no. 3, 2013.

[7] Y. Rohanachandra, D. Dahanayake, L. Rohanachandra, and G. Wijetunge, "Knowledge about diagnostic features and comorbidities of childhood autism among doctors in a tertiary care hospital," *Sri Lanka Journal of Child Health*, vol. 46, no. 1, pp. 29–32, 2017.

[8] H. Tager-Flusberg, R. Paul, C. Lord *et al.*, "Language and communication in autism," *Handbook of autism and pervasive developmental disorders*, vol. 1, pp. 335–364, 2005.

[9] A. M. Wetherby, J. Woods, L. Allen, J. Cleary, H. Dickinson, and C. Lord, "Early indicators of autism spectrum disorders in the second year of life," *Journal of autism and developmental disorders*, vol. 34, no. 5, pp. 473–493, 2004.

[10] E. L. Wodka, P. Mathy, and L. Kalb, "Predictors of phrase and fluent speech in children with autism and severe language delay," *Pediatrics*, pp. peds–2012, 2013.

[11] J. Boucher, "Language development in autism," in *International Congress Series*, vol. 1254. Elsevier, 2003, pp. 247–253.

[12] T. Charman, S. Baron-Cohen, J. Swettenham, G. Baird, A. Drew, and A. Cox, "Predicting language outcome in infants with autism and pervasive developmental disorder," *International Journal of Language & Communication Disorders*, vol. 38, no. 3, pp. 265–285, 2003.

[13] S. T. Kover, A. S. McDuffie, R. J. Hagerman, and L. Abbeduto, "Receptive vocabulary in boys with autism spectrum disorder: Cross-sectional developmental trajectories," *Journal of Autism and Developmental Disorders*, vol. 43, no. 11, pp. 2696–2709, 2013.

[14] D. K. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[15] R. Cotterell, S. J. Mielke, J. Eisner, and B. Roark, "Are all languages equally hard to language-model?" *arXiv preprint arXiv:1806.03743*, 2018.

[16] S. M. Fosnot and S. Jun, "Prosodic characteristics in children with stuttering or autism during reading and imitation," in *Proceedings of the 14th international congress of phonetic sciences*, 1999, pp. 1925–1928.

[17] M. Sharda, T. P. Subhadra, S. Sahay, C. Nagaraja, L. Singh, R. Mishra, A. Sen, N. Singhal, D. Erickson, and N. C. Singh, "Sounds of melody-pitch patterns of speech in autism," *Neuroscience letters*, vol. 478, no. 1, pp. 42–45, 2010.

[18] K. Pek, T. Arai, and N. Kanedera, "Voice activity detection in noise using modulation spectrum of speech: Investigation of speech frequency and modulation frequency ranges," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 33–44, 2012.

[19] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[20] P. M. Chauhan and N. P. Desai, "Mel frequency cepstral coefficients (MFCC) based speaker identification in noisy environment using wiener filter," in *Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on*. IEEE, 2014, pp. 1–5.