A DEEP LEARNING BASED EVALUATION OF ARTICULATION DISORDER AND LEARNING ASSISTIVE SYSTEM FOR AUTISTIC CHILDREN

Leena G Pillai and Elizabeth Sherly

Research Associate, IIITM-K, Trivandrum, India

Professor, IIITM-K, Trivandrum, India

ABSTRACT

Automatic Speech Recognition (ASR) is a thriving technology, which permits an application to detect the spoken words or the speaker using computer. This work isfocused on speech recognition for children with Asperger's Syndrome who belongs to the age group of five to ten. Asperger's Syndrome (AS) is one of the spectrum disorders in Autism Spectrum Disorder (ASD) in which most of affected individuals have high levelof IQ and sound language skills. But thespectrum disorder restricts for communication and social interaction. The implementation of this workaccomplished through two different stages. In the first stage, an attempt is made to develop an Automatic Speech Recognition system, to identify the articulation disorder inchildren with Aspergers' Syndrome (AS) on the basis of 14 Malayalam vowels. A deep auto-encoder is used for pre-training in an unsupervised manner with 27,300 features that includes Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing, Formants 1, Formants 2 and Formants 3. An additional layer is added and fine-tuned to predict the target vowel sounds, which enhance the performance of the auto–encoder. In the second stage, Learning Assistive System for Autistic Child (LASAC)is envisioned to provide speech training as well as impaired speech analysis for Malayalam vowels. A feature vector is used to create the acoustic model and the Euclidean Distance formula is used to analyze the percentage of impairment in AS student's speech by comparing their speech features against the acoustic model.LASAC provides an interactive interface that offers speech training, speech analysis and articulation tutorial. Most of the autistic students have difficulty to memorize words or letters but they are able to recollect pictures. This serves as a learning assistive system for Malayalam letters with their corresponding images and words. The DNN with auto-encoder has achieved an average accuracy of 65% with the impaired (autistic) dataset.

KEYWORDS

Automatic Speech Recognition, MFCC, Zero crossing, Formants analysis, Autism Spectrum Disorder, Deep Auto Encoder, Euclidean Distance.

1. INTRODUCTION

The speech communication is a speech chain communication that consists of speech production mechanism in the speaker, transmission through a medium and speech perception process in the ear and brain of the listener.The Automatic Speech Recognition (ASR) allows users to communicate with the computer interface,by using their voice, in a way that resembles normal human conversation [1]. In ASR, algorithm plays the listener's role by decoding speech and the speech synthesis performs the speaker's role.This paper is focused on ASR that identifies the impaired vowel voice articulated by the children with Aspergers' Syndrome.Sounds are the outcome of the articulation process, where the air coming from the vocal tract is adjusted by lips, tongue, jaw, teeth, and palate. An articulation problem leads a person to produce sounds incorrectly. Most of the autistic children deliver impaired speech. Children with articulation disorder may delete, substitute, add or distort sounds. It is observed that students get trained with sounds in their early age can reduce their articulation problem, language and communication

disorder [2].The parts of brain, associated together in the process of speech are Frontal Lobe, Parietal Lobe, Occipital Lobe and temporal Lobe. While the human being communicating, these four lobes are coupled together. In autistic children, lack of information sharing occurs in between the brain lobes. This is one of the situations that lead to communication disorder among the autistic children.

Aspergers' Syndrome (AS) is one of the Autism Spectrum Disorders (ASD), which is characterized by a high functioning form of autism with serious deficiencies in social and communication skills [6]. The IQ levels of these students vary from typically normal to very superior range [7]. Usually Asperger syndrome affected individuals have sound language skills, but they may struggle to realize the anticipations of others within conversations, or follow echolalia (repeating what the other person has just said). AS cannot be cured, but the syndrome can be reduced rehabilitative services. Language and speech therapy is one of the rehabilitative services. In this present work LASAC envisioned to provide speech training in Malayalam language as well as impaired speech and articulation analysis with the AS affected children who belong to the age group of five to ten.

This work initially concentrated on developing speech recognition for such impaired ASD children. The study was performed on articulation and voice disorder to identify the suitable feature extraction and prepared a training dataset for impaired speech recognition. The feature vector selected for network training is an accelerating factor of classification accuracy. Frequency describes the fundamental characteristics of speech signals, therefore frequency based feature extraction methods Mel-Frequency Cepstral Coefficient (MFCC) and Spectrogram analysis are used for frequency feature extraction [11,12,13]. The spectrogram analysis identifies Formants values that correlated with different articulation position. As most of the vowels are articulated by unrestricted airflow, a time based feature extraction method zero crossing used to measure the zero crossing rates in the signal. The feature vector consist of 27,300 features that extracted usingfrequency based feature extraction methods (MFCC and Spectrogram analysis)and time based feature extraction method (Zero Crossing).

To evaluate the articulation and voice disorder in children with AS, an articulatory phonetics conducted by applying unsupervised Deep Neural Network with auto-encoder. Hinton et al., (2006) initially proposed a scheme with greedy, unsupervised and layer-wise pre-training algorithm in Deep Belief Network (DBN) where each layer modeled by a Restricted Boltzmann Machine (RBM) [2]. Later works proved that architectures like auto-encoders [3] or conventional neural networks [4] with similar scheme are suitable for building Deep Networks. DNN is a type of ANN that contains multiple hidden layers between the input and output layers [5]. Complex non-linear relationships can be modeled by DNN. In this work, an auto-encoder is used for pre-training in an unsupervised manner, not to produce classification at output instead it reproduces input at output layer. Auto-encoder helps to identify relevant features by pre-training and setting the target values to be equal to the input.

In the second part, an assistive learning system called LASAC is introducedthat serves an interactive Malayalam language learning assistive system for children with AS.Several comparative studies between traditional teaching methods and digital teaching methods have found that computer assisted programs were more effective in improving language as well as cognitive outcome measures of autistic children [8,9]. These studies also reveal that with the help of computer assisted and game like technologies, such children shows greater improvements in motivation and attentiveness [10]. Following these facts, a Malayalam language Learning Assistive System for Autistic Child (LASAC) is developed for limited words. It offers an interactive interface with words and corresponding images for each vowel, articulation tutorial, voice recorder and voice analysis tool. The Euclidean Distance formula is used to analyse the impairment percentage of the vowel articulated by the user.

1.1 Malayalam Vowels

The modern Malayalam alphabets consist of 15 vowel letters. These vowel sounds are classified into Monophthongs that shows one vowel quality, for exampleഅ (a) and Diphthongs that show two vowel quality, for example ഔ (au).Malayalam is one of the richest languages in terms of number of alphabets and the toughest language while considering speech. Vowel sounds are articulated due to unrestricted airflow.Therefore, these phonemes have highest intensity and duration lies in between 60 to 400ms in normal speech. The vowel sound signals are quasi-periodic due to repeated excitation of the vocal tract.
The Malayalam vowels are classified based on tongue height, tongue backness, lip rounding tenseness of the articulators [3].

1.1.1 Tongue Height

Vowels are classified into high, low and mid based on the space between the tongue and the roof of the mouth while producing sound. The high vowelsഇ (i), ഈ (i:), ഉ (u)andഊ (u:) requires relatively narrow space between the tongue and the roof of the mouth. The low vowelsഅ (a)andആ(a:) requires relatively wide space between the tongue and the roof of the mouth. The mid vowelsഎ(e), ഏ(e:), ഒ(o) andഓ(o:) requires a space between the low and high.

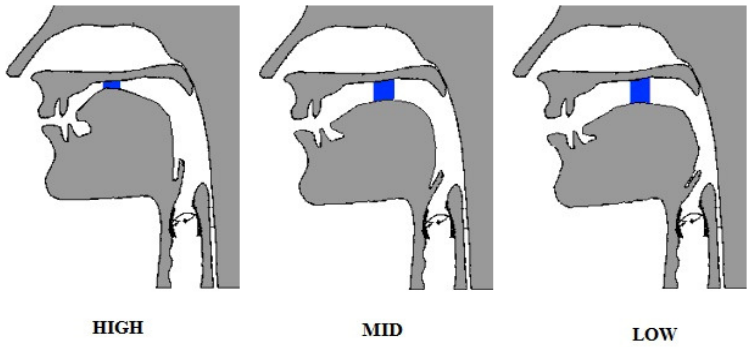

Figure 1. Tongue height position

1.1.2 Tongue Backness

Vowels are classified into three- Front, Back and Central- based on how far the tongue is positioned from the back of the mouth. The front vowelsഇ(i), ഈ(i:), എ(e) andഏ(e:)are articulated by placing the tongue forward in the mouth. The back vowelsഉ(u), ഊ(u:), ഒ(o) andഓ(o:)are articulated by placing the tongue far back in the mouth. The central vowelsഅ (a), ആ (a:)are articulated by placing the tongue roughly in between the front and the back of the mouth.

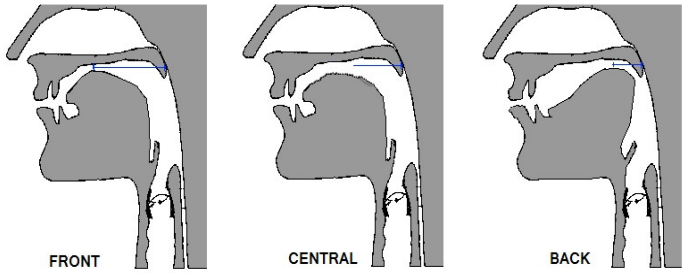

Figure 2. Tongue backness position

1.1.3 Lip Rounding

Another aspect of vowel classification is based on the presence or absence of lip rounding while producing a sound. Vowels -ഉ(u), ഊ(u:), ഒ(o), ഓ(o:) andഔ (au) are produced with a high degree of lip rounding.

		Front	**Central**	**Back**
High	Short	**ഇ i**		**ഉ u**
	Long	**ഈ i:**		**ഊ u:**
Mid	Short	**എ e**		**ഒ o**
	Long	**ഏ e:**		**ഓ o:**
Low	Short		**അ a**	
	Long		**ആ a:**	

Table 1.Malayalam Vowel Classification.

2. LITERATURE REVIEW

Manasi Dixit and ShailaApte (2013) conducted a study on speech evaluation with children suffering from Apraxia of Speech (AOS). The work is based on the study of children in the age group of three to eight with Marathi mother tongue [15]. The speech disability of AOS students, estimated based on the result retrieved from Speech Pause Index, Jitter, Skew and Kurtosis analysis.The Speech-Pause Index (SPI) that indicates Segmental and Supra-segmental analysis was remarked to be high in pathological subjects. The skew factor identified below the threshold level of 3.4 and the Kurtosis factor identified below 12 in almost all pathological subjects. This work observed that formants are widely spread in case of pathological subjects.

Nidhyananthanet al., (2014) conducted an experiment on contemporary speech or speaker recognition with speech from impaired vocal apparatus [16]. This work performed a comparative study between the MFCC, ZCPA and LPC feature extraction algorithm and modelling methods GMM, HMM, GFM and ANN. These feature extraction algorithms and modelling methods applied on speech recognition of normal speakers and persons with speech disorders. The MFCC, ZCPA and LPC feature extraction techniques are compared and analysed. Similarly GMM, HMM, GFM, ANN and their fusions are analysed and evaluated with their identification accuracy. This work concluded that ASR system gives high performance with normal speakers but in case of persons with speech disorders it gives low performance. This work suggests that in order to improve the accuracy, multi nets ANN can be employed. MFCC feature extraction has acquired the highest accuracy rate of 85%, whereas ZCPA has acquired 38% and LPC has achieved 82%. Ahmed et al., (2012)conducted an ASR experiment in Malay language with MFCC features. This work has obtained an average accuracy of 95% using Multilayer Perceptron [17].

Kyung-Im Han, Hye-Jung Park and Kun-Min Lee (2016) conducted a study on hearing impaired people who mostly depends on visual aids [18]. SVM technique used to extract phonetic features

of the sound and lip shapes was examined for each English vowel. This work allows the language learners to see a visual aid that provides guidelines regarding the exact lip shape, while producing a vowel. F1 and F2 formants frequency analysis enhance for the identification of lip shape. This paper implemented the lip shape refinement for five English vowels. The speakers involved in this work are English native speakers, Korean normal-hearing speakers and Korean hearing impaired speakers. In this work, it is observed that lip shape adjustment is very important to acquire native-like sounds.

Anu V Anand, P Sobhana Devi, Jose Stephen and Bhadran V K (2012) were applied Malayalam Speech Recognition System for visually challenged people [19]. MFCC method used for feature extraction and Hidden Markov Model (HMM) used to build Acoustic model. For this study, samples collected from 80 speakers. In this work, ASR with Speech-To-Text (STT) in openoffice writer converts the visually challenged individual's speech to text and Text-To-Speech (TTS), reads the executed commands and documents. This system works in two mode dictation mode and command mode. It offers user friendly interface for visually challenged people for easy document preparation.

Thasleema et al., (2007) have conducted a K-NN pattern classifier based on Malayalam vowel speech recognition. Linear Predictive Coding (LPC) method is used for Malayalam vowel feature extraction[20]. The result compared with wavelet packet decomposition method. The k-NN pattern classifier used for Recognition. The study conducted on five Malayalam vowels അ (a), ഇ (i), ഏ (e), ഒ (o) and ഉ (u) only. The overall recognition accuracy obtained in five Malayalam vowels using wavelet packet decomposition method is 74% whereas the recognition accuracy obtained by using LPC method is 94%.

Several studies with ASD have been proved that the students with autism spectrum disorder responds well to treatments or therapies that involve visual support such as pictures [21], videos [22] and computers [23]. The picture exchange communication system (PECS) is an augmentative communicationsystem frequently used with children with autism.The introduction of the video modelling intervention led to an increase in both verbal and motor play responses in the children with autism.

3. METHODOLOGY

The present work is conducted with articulatory phonetics on 14 Malayalam vowels to identify the voice production impairment of Autism affected children as well as a learning assistive system for corrective measurement.

For pre-processing, the spectral noise gate technique is applied for noise reduction using audacity.The feature extraction methods that are employed in this work are MFCC, Zero-crossing and spectrogram analysis. Thenetwork is trained using Deep Neural Network with auto-encoder for the evaluation of articulation disorder. In the learning assistive interface, Euclidean Distance is used to analyse the level of impairment in autistic speech.

3.1 Mel-Frequency Cepstral Coefficients (MFCC)

The articulated phoneme can be predicted accurately by identifying the shape of the vocal tract because the shape of the vocal tract determines the variations in sound wave. This shape is enveloped with in the short term power spectrum. The MFCC feature extraction algorithm extract features from the short term power spectrum. Therefore, Mel-Frequency Cepstral Coefficient (MFCC) considered as a most commonly used faster and prevailing feature extraction method in ASR.The MFCC comprises of eight steps that depicted in the Figure3.The pre-emphasis stage is employed to boost the amount of energy in the high frequencies so that the acoustic model can be

formed with quality information. The signals are enclosed in to 30ms with an overlap of 20ms frames. The short frames are considered with the assumption that the audio signals do not change much in short time scale.

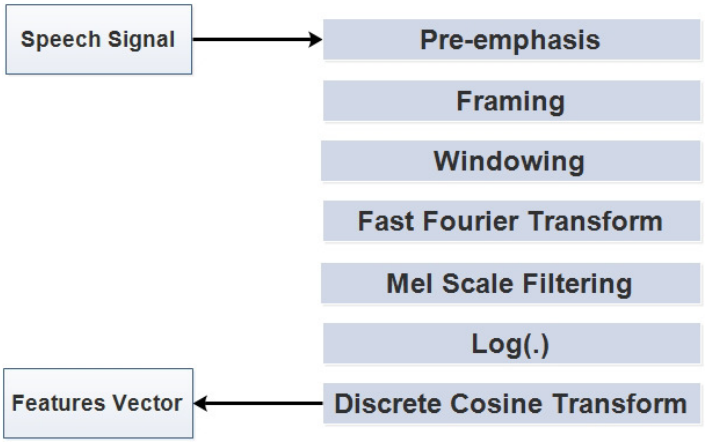

Figure 3. MFCC Workflow

The power spectrum calculation of each frame is inspired by cochlea (an organ in the ear) that vibrates at different spots depending on the frequency of the incoming sounds. Depending on this vibration nerves fires information to the brain about the presence of frequencies. The MFCC also identifies the frequencies that present in the frame. The Filter bank concept is also adopted from cochlea, which identifiesenergythat exists in various frequency regions.

3.2 ZERO – CROSSING RATE

Zero crossing is defined by measuring the number of times the amplitude of the speech signal crosses a value of zero within a given window [25]. According to the presence or absence of the vibration in the vocal cord, sounds can be classified in to voiced or unvoiced. Voiced speech produced as a result of the vibration of the closed vocal tract by the periodic air pressure at the glottis and this segment shows a low zero crossing count. Unvoiced speech produced due to the free air flow in the open vocal tract and shows a high zero crossing rates (Figure 4). The vowel sounds are produced as a result of open and unrestricted air flow from the vocal tract.

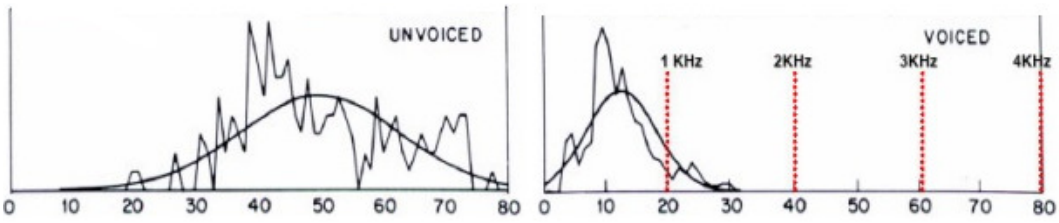

Figure 4. Zero crossing rate in unvoiced and voiced

3.3 SPECTROGRAM ANALYSIS

A visual representation that displays the spectral data over time with the amplitude of the frequency components denoted by different colours or usually different shades of gray is called Spectrogram. In spectrogram the horizontal axis represents the duration of sound in second or

millisecond. The vertical axis shows the frequency values. The most important formants parameter can be identified by the means of degree of darkness in the spectrogram [13].

Vowels are fully developed formants pattern that can be classified on the basis of the F1 and F2 analysis. These resonance frequencies associated with the two cavities separated by the constriction in the vocal tract. F1 is the lowest frequency associated with tongue height that affects space of the pharynx. F2 is the second lowest frequency associated with the changes within the oral cavity correlated with lip, jaw or tongue retraction.

For example the vowel ഇ (i) is produced from the front constriction with large pharyngeal cavity and small oral cavity. Therefore ഇ (i)can be characterized by low F1 around 400Hz and high F2 around 2500Hz. The vowel അ (a) is produced from the back constriction with much smaller pharyngeal cavity and large oral cavity. Therefore അ (a)can be characterized by slightly high F1 around 600Hz and much low F2 around 1000Hz (Figure 5). F1 and F2 values are enough for vowel classification, but for improving accuracy and reliability F3 value is also considered. The F3 value analysis helps to make a differentiation between rounded and unrounded vowels.

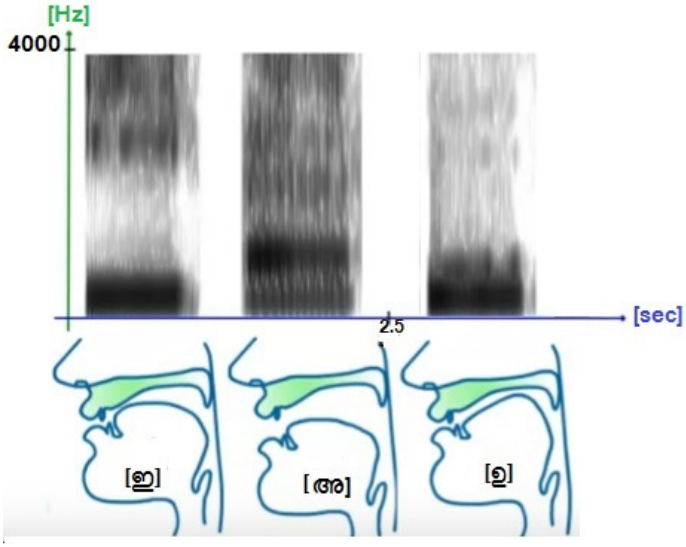

Figure 5: Spectrogram representation of അ (a), ഇ (i) and ഉ (u)

3.5 DEEP NEURAL NETWORK

DNN is a class of ANN that composed of multiple layers with completely connected weights and initialized by applying unsupervised or supervised pre-training techniques for solving classification problems in complex data [5]. The nodes in each layer trained with distinct level of feature abstraction or feature hierarchy, which enables a DNN to solve very large, high-dimensional dataset with large set of parameters [3].

In this work, the unsupervised DNNs are effective method to discover latent structures and patterns with in unlabelled and unstructured. Each node in a DNN learns structures automatically from the unlabelled data by reconstructing the size and dimension of the input. Therefore, the network generates correlations between convinced features and optimal results. The individual layer-wise training can be achieved by a special type of nonlinear method called auto-encoder without the class labels [4]. Deep-learning ends in a labelled softmaxoutput layer that joins all the hidden layers together to form a deep network and train in a supervised fashion.

4. IMPLEMENTATION

4.2 DATA COLLECTION AND PRE-PROCESSING

The training data set have been collected from 30 normal students, 13 boys and 17 girls. For testing, vowel sounds are recorded from 10 students, 3 boys and 7 girls, with AS. The ZoomH1 recorder is used for this recording purpose. The records are collected in .wav format in 48 KHz sampling rate with 24-bit resolution.

The Undesired frequencies that appear in the speech signal that degrades the quality of signalsare considered as noise. The spectral noise gate technique is applied for noise reduction using audacity. The decibel passed to Noise Reductionis 12 dB to reduce the noise to an acceptable level. The Frequency smoothing set to 3Hz for speech smoothening.

4.3 FEATURE EXTRACTION

The three algorithmicfeature extraction Procedures that are implemented in this paper are described in Table 2.

Procedure 1

Input : Speech Signals
Output : 60 MFCC features

1. Boost the amount of energy in the high frequencies (Pre-emphasis).
2. Signal is divided in to 30ms frames with an overlap of 20ms& and apply spectral analysis.
3. Each frame multiplied by window function y[n] = x[n] * w[n].
4. Apply FFT to each frame to convert the samples from time domain tofrequency domain.

$$X_k = \sum_{x_n}^{N-1} \left(cos\left(-2\pi k \frac{n}{N}\right) + i.sin\left(-2\pi k \frac{n}{N}\right)\right)$$

5. A group of triangle band pass filters that simulates the characteristicsof the human's ear are applied to the spectrum.
6. Compute the logarithm of square magnitude of the output of Mel-filter bank.

$$m = 2595 \log 10\left(1 + \left(\frac{f}{100}\right)\right)$$

7. DCT convert the log Mel spectrum in to time domain using Discrete Cosine.

$$X_n = \frac{1}{N} \sum_{k=0}^{N-1} N - 1\, x_k \left(cos\left(2\pi k \frac{n}{N}\right) + i.sin\left(2\pi k \frac{n}{N}\right)\right)$$

8. Return 60 MFCC values.

Procedure 2:

Input : Speech Signals
Output : Zero Crossing Rate and Standard Deviation

1. Set Sampling_rate=48000

2. Calculate the length of the window

$$\text{Window\_length} = \text{Frame\_size X Sampling\_rate}$$

3. Calculate the number of frames

$$\text{Frame\_num} = (\text{length(wave\_file)} - \text{Window\_length} / \text{Sampling\_rate}) + 1$$

4. In each frame
   Calculate Zero Crossing Rate

$$ZCR = \frac{1}{2N}\sum_{n=1}^{N}|sign(x[n]) - sign(x[n-1]|$$

   Calculate Zero Crossing Standard deviation

$$std = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(ZCR_i - \mu)^2}$$

5. Return ZCR and std

Procedure 3:

Input : Speech Signals, TextGrid
Output : Formants 1, Formants 1, Formants 1

1. Set Interval=3
2. Open each sound file and corresponding TextGrid

   Get the label of intervals with number of tiers and number of intervals
   Label = Get label of interval : tiernum, intnum

Calculate the Midpoint
   Beg = Get Start point : tiernum, intnum
   End = Get End point : tiernum, intnum
   Midpoint = Beg + ((End – Beg)/2)

   Calculate formants values at each interval
   F1 = Get peak frequency at time 1 midpoint hertz
   F2 = Get peak frequency at time 2 midpoint hertz
   F3 = Get peak frequency at time 3 midpoint hertz

3. Return F1, F2 and F3 from each file

Table 2: Algorithmic procedures for feature extraction

The feature extraction algorithms created a feature vector from the processed signals. The features that are extracted from a single vowelare listed in Table 3.

Feature Extraction	Number of features
MFCC	60
Zero Crossing	2
Spectrogram Analysis	3
Total	**65**

Table 3: Features extracted from a single vowel

To create the feature vector records collected from 30. 14 vowels are sampled from each dataset (30 X 14 = 420) and 65 features retrieved from each sample. Therefore, the feature vector consist of 27,300 (420 X 65) features.

4.3 AUTO – ENCODER

An auto-encoder is a greedy layer-wise approach for pre-training that consists of an encoder and a decoder. The encoder is responsible to map the input to a hidden representation and the decoder map this representation back to the original input (Figure 7).

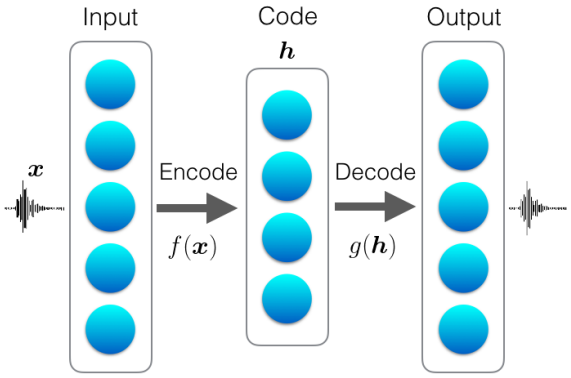

Figure 7: Deep Auto – encoder Architecture

An auto-encoder consists of multiple layers of sparse encoders. Each layer outputs are wired to the input of successive layer. Consider an auto-encoder with n number of layers. $W^{(1)}$, $W^{(2)}$, $b^{(1)}$, $b^{(2)}$are the parameters for k^{th} auto-encoder.

Then the encoding step is given by forward order:-

$$a^{(l)} = f\left(z^{(l)}\right) \tag{1}$$

$$z^{(l+1)} = W^{(l.1)}a^{(l)} + b^{(l,1)} \tag{2}$$

The decoding step is given by reverse order:-

$$a^{(n+l)} = f(z^{(n+l)}) \tag{3}$$

$$z^{(n+l+1)} = W^{(n-l,2)}a^{(n+l)} + b^{(n-l,2)} \tag{4}$$

where the information of interest is stored in $a^{(n)}$.W denotes the weight matrix and b represents bias vector. The features retrieved from the auto-encoder can be used for classification problems by passing $a^{(n)}$to a softmax classifier.In the first layer, regularizersare applied to learn the sparse representation. The control parameters given to the regularizers are listed in Table 4.

Hidden size	
Auto-encoder 1	65
Auto-encoder 2	60
L2WeightRegularization	0.002
SparsityRegularization	4
SparsityProportion	0.02
DecoderTransferFunction	Purelin

Table 4. Auto-encoder parameters

Two auto-encoders are implemented in this work. 65 coefficients (features extracted from each sound samples) are given to the first auto-encoder. The number of neurons in hidden layer is 65 (equal to the number of features), therefore the first auto-encoder delivers output of 65 X 1 dimensions. The extracted features from the first encoder are given as an input to the second auto-encoder. The number of neurons in second hidden layer is set to 60 which results a compressed output with 60 X 1 dimensions (Figure 8). These features are considered as the bottle neck features for classification.

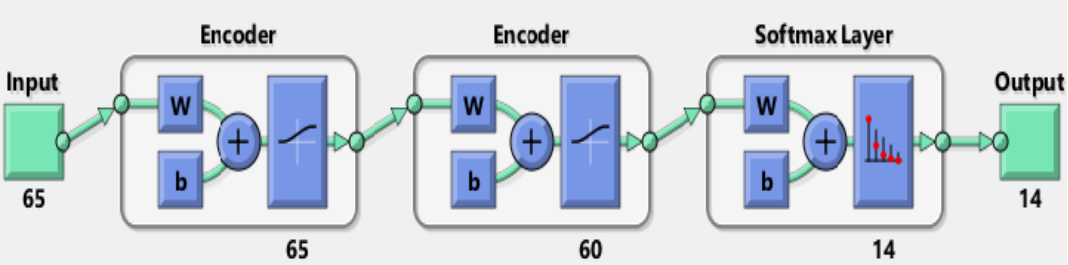

Figure 8. Implementation architecture of Auto-encoder

Finally train the softmax layer in a supervised fashion to transform all the net activations to a series of values that can be interpreted as probabilities. The softmax layer trained with dataset and the target output. The softmax layer applies softmax function or normalized exponential to the input and classifies to 14 classes.

5. RESULT AND DISCUSSION

The experiments are conductedto study on impaired speech and articulation disorder in children with Asperger's Syndromewithin the age limit of five to ten. As an initial work, this study is focused on 14 Malayalam vowels for 50 children, which includes normal and AS affected. The normal audio is included to identify the difference in AS children articulation. The accuracy is measured using true positive, true negative, false positive and false negative rates.

$$Accuracy = \frac{\#\ correctly\ predicted}{Total\ testdata} X\ 100 \quad (5)$$

$$Correctly\ predicted\ data = TP+TN$$
$$Total\ test\ data = TP+TN+FP+FN$$

TP - True Positive, TN - True Negative, FP - False Positive, FN - False negative.

5.1 AUTISTIC ARTICULATORY PHONETICS

Articulatory phonetics conducted to evaluate the articulation disorder in children with AS.This would help the learning assistive system to ascertain the articulation disorder faced by the autistic children.The Network trained with unsupervised Deep Auto-encoder.

NETWORK TRAINING

From the 14 Malayalam vowel sounds collected from both AS affected students (20 speakers) and normal (30 speakers) students. 65 features extracted from each sample.

The autistic speech datasets are able to achieve only 54% of classification accuracy whereas normal student's speech dataset obtained 98%. Each autistic child faces unique articulation and voice impairment. Therefore, identifying a classification structure from their feature vector is

difficult. The experiment reveals that records collected from normal students are preferable to train the network.

NETWORK TESTING

The Network is initially trained with an unsupervised Deep Auto-encoder. The DNN with auto-encoder is able to achieve 56% of test accuracy with the impaired (autistic) dataset. Each class obtained variant accuracy ranging from 30% to 90%. This variation is based on articulation requirement of each vowel. Table 5 describes the detailed evaluation result based on each class and the accuracy obtained from unsupervised DNN.

CLASS	LETTER	DNNACCURACY %
1	അ (a)	90
2	ആ (a:)	80
3	ഇ (i)	70
4	ഈ (i:)	70
5	ഉ (u)	50
6	ഊ (u:)	50
7	ഋ (rr)	30
8	എ (e)	70
9	ഏ (e:)	50
10	ഐ(ai)	40
11	ഒ (o)	50
12	ഓ (o:)	40
13	ഔ(au)	30
14	അം(am)	70
Total		56%

Table 5: Test result of each vowel articulated by autistic children

The performance evaluation table (Table 5) and graph (Figure 9) explicates that the children with AS struggle to produce tongue back and high or middle monophthong vowels - ഉ (u), ഊ (u:), ഒ (o) andഓ (o:). All the primary letters such as അ (a), ഇ (i), ഉ (u),എ (e) and ഒ(o)are able to achieve more accuracy than its extended letters such as ആ (a:), ഈ (i:), ഊ (u:), ഏ (e:) and ഓ (o:). Most of the autistic children are not able to identify the difference between primary and its extended letters. They articulate the same manner for both sounds.Misclassification occurs regularly in between the tongue backness featured and lip rounded vowels ഉ (u), ഊ (u:), ഒ (o) and ഓ (o:).

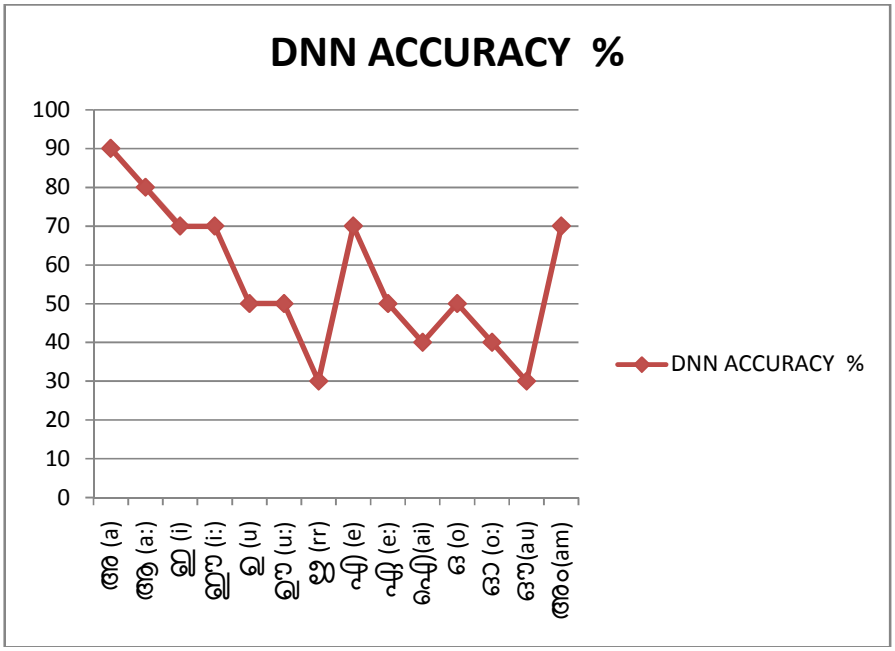

Figure 9: Performance evaluation graph of each vowel articulated by autistic children

The vowels ഐ (ai) and ഔ (au) are diphthong vowels. A diphthong vowel shows two vowel qualities. The vowel ഐ (ai) is the combination of two vowel sounds അ (a) and ഇ (i).The vowel ഔ (au)is the combination of two vowel sounds അ (a) and ഉ (u). Diphthongs vowels acquired least accuracy as they misclassified with any of their combination vowels. Most of the AS affected students are not able to produce both sounds together. The vowel sound ൠ (rr) is the combination of 'eee-rrr'. The ൠ (rr) can be produced by restricted airflow that cause the vocal cord to vibrate. To shape this sound the tip of the tongue placed in between the hard and soft palate. Therefore ൠ (rr) also attained least accuracy. Most of the autistic students are not able to articulate these (ഐ (ai), ഔ (au),ൠ (rr) and ഓ (o:)) letters properly. Therefore, as far as the DNN network is considered, without these vowels thesystem performance shows 65% of accuracy(Table 6 and Figure 10).

DATASET	**ACCURACY**
Students with AS	65%
Normal students	98%

Table 6.Performance of the Classifier

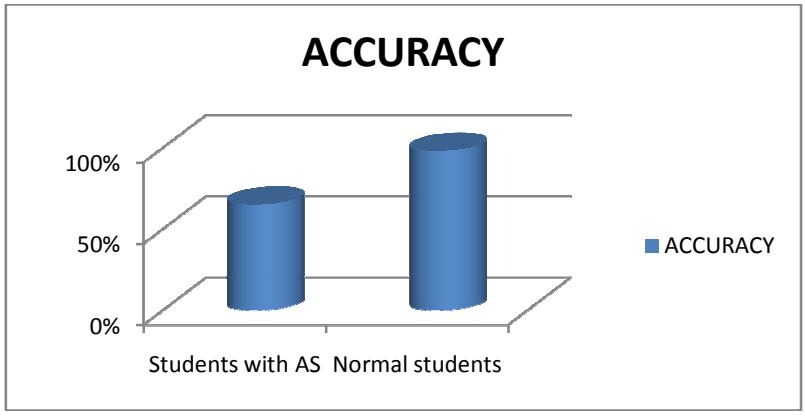

Figure 10.Autistic Vs Normal Speech Accuracy Graph

5.2 LEARNING ASSISTIVE SYSTEM FOR AUTISTIC CHILDREN (LASAC)

The LASAC provides speech training through an interactive interface, by considering the common interest of the autistic children. The interface consists of Malayalam virtual keyboard, words and corresponding images for each vowel, Articulation tutorial for each vowel, voice recording and analysis tool.The DNN classifier is used to provide assistance to the Speech Language Pathologist (SPL) in impaired speech recognition. The DNN classifies the autistic speech input in to the most suitable class according to their observed similarities. Apart from classification, the Euclidean Distance formula is used to analyse the impairment percentage in autistic voice input. The result obtained from Euclidean Distance stored in database for further evaluation and tracking the improvement of the student speech after training. The architecture of LASAC is depicted in Figure 11.

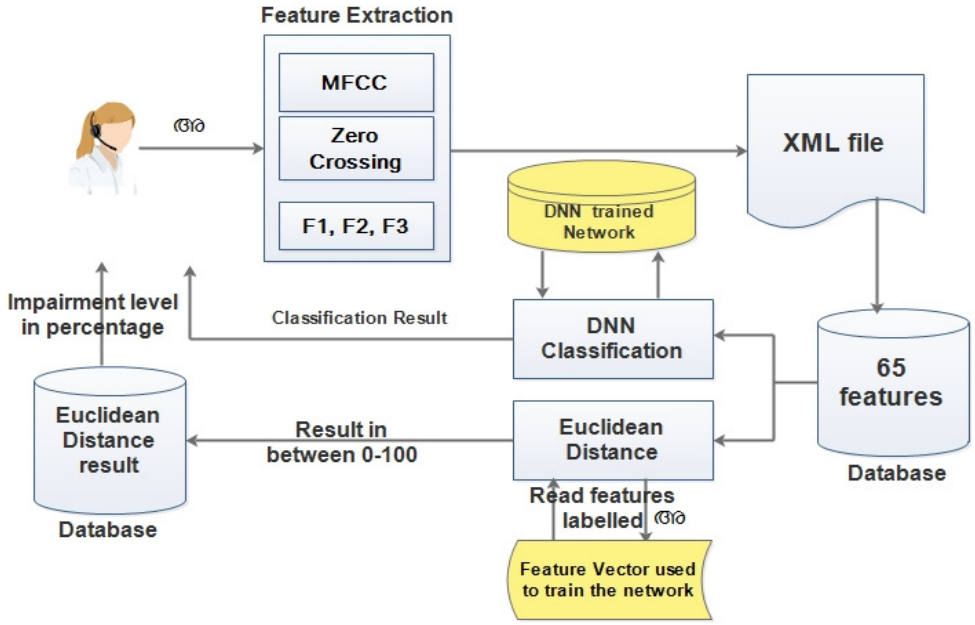

Figure 11. LASAC architecture

The straight-line distance between two points in Euclidean space is called Euclidean distance. The Euclidean Distance is the most common formula to calculate the difference between two points [26]. It calculates the root square difference between coordinates of a pair of objects. The distance formula is derived from the Pythagorean Theorem. The Euclidean Distance formula is :

$$Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

This work provides an interactive learning assistive system with speech training for the autistic children. To enhance the speech training, Euclidean Distance formula is implemented to calculate the difference between normal speech dataset and autistic speech test dataset.Figure 12 shows the GUI result obtained after Euclidean calculation.

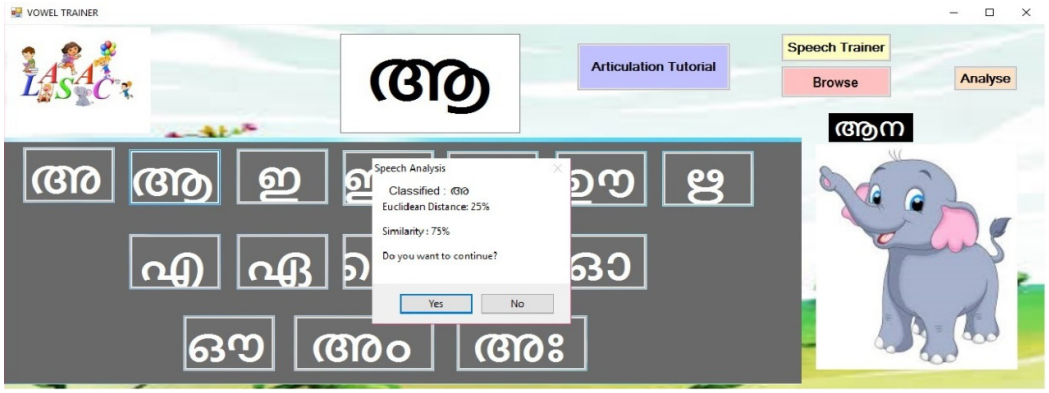

Figure 12: GUI result of Euclidean Distance

The result describes that the given voice sample is 25% different from the feature vector labelled ആ (a:), which means the sample is 75% similar to the normal ആ (a:) and the DNN classifier classified the input voice to അ (a). The result is recorded in database for future analysis.

Phoneme	Similarity	Remark
അ (a)	90%	Good
ആ (a:)	75%	Average
ഇ (i)	80%	Average
ഈ (i:)	72%	Average
ഉ (u)	66%	Below Average
ഊ (u:)	62%	Poor
ഋ (rr)	51%	Poor
എ (e)	83%	Average
ഏ (e:)	76%	Average
ഐ(ai)	60%	Poor
ഒ (o)	64%	Below Average
ഓ (o:)	59%	Poor
ഔ(au)	53%	Poor
അം(am)	85%	Good

Table 7. Similarity list achieved from an autistic student's speech training

Table 7 shows the Similarity list achieved from speech training, provided to an autistic student and the corresponding remarks. This result helps to analyze the basic speech impairment percentage level of the AS students. Guardians or teachers can focus on those vowels for further training which shows least similarity.

6. CONCLUSION

This work is coupled with evaluation of articulation disorder in AS students and Malayalam Vowel training as well as impaired voice analysis. The DNN with auto-encoder, an unsupervised machine learning technology, used to analyse the articulation disorder in the AS students in the age group of five to ten. Two auto-encoders are used with 65 and 60 hidden layer neurons respectively, which enhanced to achieve improved result with bottle neck feature extraction.This work is based on 14 Malayalam vowels.Though the network trained with 27,300 features has shown an average classification accuracy of 98% for normal children, the Autism children's accuracy obtained is only 65%.

An interactive interface used to provide language and speech training. In speech training, the DNN classifier used to assist the Speech Language Pathologist (SPL) to recognize the autistic speech input and Euclidean Distance formula is used to enhance the speech training by evaluating thepercentage of impairment. The feature vector with 65 features contains MFCC (60), Zero Crossing Rate (2) and Spectrogram analysis (3).

The Centers for Disease Control and Prevention (CDC) reveals that 1 among the 68 children are diagnosed with Autism Spectrum Disorder (ASD).In this scenario, the technologies required to support the communication and language development of these children.In future, this work can be extended to consonants identification along with more unique features. The maximum features can be combined and with the help of bottle neck feature extraction it is able to identify optimum features that leads to a better performance. As this work is social relevant, we hope that this work will inspire others to conduct more research on this area that provide support to the communication and language development of autistic children.

REFERENCES

[1] Atal, B., and L. Rabiner."A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition."IEEE Transactions on Acoustics, Speech, and Signal Processing 24, no. 3 (1976): 201-212.

[2] G.E. Hinton, S. Osindero, and Y.W. Teh, "A fast learning algorithm for deep belief nets," Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[3] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," Advances in neural information processing systems, vol. 19, pp. 153, 2007.

[4] M.A.Ranzato, F.J. Huang, Y.L. Boureau, and Y. LeCun. "Unsupervised learning of invariant feature hierarchies with applications to object recognition."Computer Vision and Pattern Recognition, 2007.CVPR'07.IEEE Conference on. IEEE,2007, pp. 1–8.

[5] Deng, Li, Geoffrey Hinton, and Brian 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, pp. 8599-8603. IEEE, 2013.

[6] Bowler, Dermot M. ""Theory of Mind" in Asperger's Syndrome Dermot M. Bowler." Journal of Child Psychology and Psychiatry 33, no. 5 (1992): 877-893.

[7] Ozonoff, Sally, Sally J. Rogers, and Bruce F. Pennington. "Asperger's syndrome: Evidence of an empirical distinction from high-functioning autism." Journal of Child Psychology and Psychiatry 32, no. 7 (1991): 1107-1122.

[8] Ozonoff, Sally. "Reliability and validity of the Wisconsin Card Sorting Test in studies of autism." Neuropsychology 9, no. 4 (1995): 491.

[9] Whalen, Christina, Debbie Moss, Aaron B. Ilan, ManyaVaupel, Paul Fielding, Kevin Macdonald, Shannon Cernich, and Jennifer Symon. "Efficacy of TeachTown: Basics computer-assisted intervention for the intensive comprehensive autism program in Los Angeles unified school district." Autism 14, no. 3 (2010): 179-197.

[10] Williams, Christine, Barry Wright, Gillian Callaghan, and Brian Coughlan. "Do children with autism learn to read more readily by computer assisted instruction or traditional book methods? A pilot study." Autism 6, no. 1 (2002): 71-91.

[11] Majeed, Sayf A., Hafizah Husain, Salina Abdul Samad, and Tariq F. Idbeaa. "Mel Frequency Cepstral Coefficients (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition: a Comparison Study." Journal of Theoretical and Applied Information Technology 79, no. 1 (2015): 38.

[12] Hossan, MdAfzal, SheerazMemon, and Mark A. Gregory. "A novel approach for MFCC feature extraction." In Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on, pp. 1-5. IEEE, 2010.

[13] Jiang, J. J., and Yu Zhang. "Nonlinear dynamic analysis of speech from pathological subjects." Electronics Letters 38, no. 6 (2002): 294-295.

[14] Atal, B., and L. Rabiner. "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition." IEEE Transactions on Acoustics, Speech, and Signal Processing 24, no. 3 (1976): 201-212.

[15] Dixit, Manasi, and ShailaApte. "Speech evaluation with special focus on children suffering from apraxia of speech."Signal and Image Processing 4, no. 3 (2013): 57.

[16] Nidhyananthan, S. Selva, R. ShanthaSelvakumari, and V. Shenbagalakshmi. "Contemporary speech/speaker recognition with speech from impaired vocal apparatus." In Communication and Network Technologies (ICCNT), 2014 International Conference on, pp. 198-202. IEEE, 2014.

[17] Ahmed, Irfan, Nasir Ahmad, Hazrat Ali, and Gulzar Ahmad. "The development of isolated words pashto automatic speech recognition system." In Automation and Computing (ICAC), 2012 18th International Conference on, pp. 1-4. IEEE, 2012.

[18] Han, Kyung-Im, Hye-Jung Park, and Kun-Min Lee. "Speech recognition and lip shape feature extraction for English vowel pronunciation of the hearing-impaired based on SVM technique." In Big Data and Smart Computing (BigComp), 2016 International Conference on, pp. 293-296. IEEE, 2016.

[19] Anand, Anu V., P. Shobana Devi, Jose Stephen, and V. K. Bhadran. "Malayalam Speech Recognition system and its application for visually impaired people." In India Conference (INDICON), 2012 Annual IEEE, pp. 619-624.IEEE, 2012.

[20] Thasleema, T. M., V. Kabeer, and N. K. Narayanan. "Malayalam vowel recognition based on linear predictive coding parameters and k-nn algorithm." Conference on Computational Intelligence and Multimedia Applications, 2007.International Conference on.Vol. 2.IEEE, 2007.

[21] Charlop-Christy, Marjorie H., Michael Carpenter, Loc Le, Linda A. LeBlanc, and Kristen Kellet. "Using the picture exchange communication system (PECS) with children with autism: Assessment of PECS acquisition, speech, social-communicative behavior, and problem behavior." Journal of applied behavior analysis 35, no. 3 (2002): 213-231.

[22] D'Ateno, Patricia, Kathleen Mangiapanello, and Bridget A. Taylor. "Using video modeling to teach complex play sequences to a preschooler with autism." Journal of Positive Behavior Interventions 5, no. 1 (2003): 5-11.

[23] Hetzroni, Orit E., and JumanTannous. "Effects of a computer-based intervention program on the communicative functions of children with autism." Journal of autism and developmental disorders 34, no. 2 (2004): 95-113.

[24] Wong, Eddie, and SridhaSridharan. "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification."In Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on, pp. 95-98. IEEE, 2001.

[25] Djurić, Milenko B., and Željko R. Djurišić. "Frequency measurement of distorted signals using Fourier and zero crossing techniques." Electric Power Systems Research 78, no. 8 (2008): 1407-1415.

[26] Wang, Liwei, Yan Zhang, and JufuFeng. "On the Euclidean distance of images."IEEE transactions on pattern analysis and machine intelligence 27, no. 8 (2005): 1334-1339.

AUTHORS

Leena G Piilai is currently working as Research Associate under Dr. Elizabeth Sherlyat Indian Institute of Information Technology andmanagement-Kerala (IIITM-K), Technopark, Trivanrum, Kerala. Has completed MPhilin Computer Science from Cochin University of Science And Technology (CUSAT), Kerala. Herresearch area is Automatic Speech Recognition and for the last one year she is focusing in speech analysis and recognition of Autistic children.

Dr. Elizabeth Sherly is currently working as Professor ofIndian Institute of Information Technology andmanagement-Kerala (IIITM-K), a Kerala Governmentorganization inTechnopark, Trivanrum, Kerala. Havingmore than 25 years of experience in Education andResearch in CS/IT, her major research work involves inPattern recognition and Data mining using Soft computing.Ph.D in Computer Science from University of Kerala inArtificial Neural Networks in Biological Control Systems,now actively pursuing research in Medical ImageProcessing, Computational Linguistics, Automatic SpeechRecognition and Object Modeling Technologies etc..