International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Neural Network-Based Approach for Identification and Classification of Speech Disfluency: The Apraxia of Speech

[1]Ashwini P., [2]S. H. Bharathi

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: Over the past decade, the field of signal processing has witnessed remarkable growth, particularly in speech processing, with a substantial impact from the integration of Artificial Intelligence (AI) and Machine Learning (ML). The focus on AI/ML-based speech processing has notably advanced in the identification of voice disfluencies, particularly within biomedical applications. Given the critical nature of disfluency identification, the range of potential applications is extensive, as these inconsistencies pose challenges to effective human communication. This paper specifically delves into the examination of apraxia of speech, presenting an algorithm designed for its identification—a distinctive form of speech disfluency. The algorithm is built upon a Convolutional Neural Network (CNN) deep neural network, forming the cornerstone of its development for categorizing normal and apraxic speech. Feature extraction involves the utilization of Teager energy operators, encompassing fundamental frequency, Mel-Frequency Cepstral Coefficients (MFCC), short-term zero crossing rate (STZCR), and Teager energy (TEO). Notably, the incorporation of STZCR as a classification parameter significantly enhances the classifier's efficiency compared to TEO. The inclusion of STZCR results in an impressive 89 percent efficiency in speech categorization, whereas TEO alone yields an efficiency of 80 percent. This research underscores the pivotal role of AI/ML-based approaches in addressing speech disfluencies, particularly in the context of apraxia, contributing to advancements in early and accurate identification of communication disorders.

Keywords: *Apraxia of speech, Voice activity, Zero crossing detection, MFCC, CNN classifier*.

1. Introduction

Apraxia of speech (AOS) is an uncommon neurological speech disorder in which an individual will not be able to accurately move his/her mouth and auditory organs as easily as a normal individual. This condition is caused when the brain has difficulty directing and coordinating with auditory organs. With this disorder, the speech muscles are not considered to be weak, but they are unable to perform normally. The individual knows exactly what he/she wants to say, but there is a disruption in the part of the brain that sends the signal to the muscle for the specific movement [1]. Individuals with acquired AOS demonstrate hallmark characteristics of articulation and prosody (rhythm, stress, or intonation) errors. [1] [2] Coexisting characteristics may include groping and effortful speech production with self-correction, difficulty initiating speech, abnormal stress, intonation and rhythm errors, and inconsistency with articulation. [3] Diagnosing and treating apraxia of speech can be done by 3 methods i) with help of speech-language pathologist-specific exams that measure oral mechanisms of speech, ii) Real-time magnetic res- onance imaging (MRI) and accompanying analytical methods iii) using an automatic speech recognition system. Early detection of AOS may reduce the risk of long-term persistence of the problem.

The initial sign that a parent often notices concerning communication is that the child exhibits difficulty to understand. The child may be receiving new words very gradually, but the kid may not be able to make sentences. The reason behind this unambiguousness and poor acquiring of new words possibly will be owing to an articulation delay or a disorder. In some cases, there may be overlapping opinions between articulation delay and developmental apraxia or apraxia of speech. The difference between both can be evaluated as follows: With both the speech disfluencies articulation delay and apraxia, the child is very intelligible or very unpredictable. But the major difference lies in an articulation delay, the child has trouble with limited sounds. This can usually be relieved fairly fast. If apraxia is involved, therapy will take quite long to help the patient to overcome it [4].

Developmental Apraxia is a motor speech disorder that disturbs the coordination and planning of sounds and sound combinations. For an appropriate pronunciation of speech sounds, the articulators (which are the lips, tongue, jaw, palate, and vocal folds) need to be finely coordinated and flow correctly ordered from one sound to the consecutive sound. Apraxia pro- duces discoordination between all the articulators. Rigorous speech therapy is presently the best method followed for recovering from this disorder [4].

[1] *Research Scholar School of ECE, REVA University, Bangalore, India*
ORCID ID : : 0000-0003-0817-6732
2 Professor, School of ECE, REVA University, Bangalore, India
ORCID ID : 0000-0002-3993-5968
** Corresponding Author Email: ashwinip@reva.edu.in*

In the proposed work CNN and some of the ML algorithms have been used to do a comparative study in the identification of Apraxia. Further, VAD is done based on two different algorithms and proposed that STZCR can effectively identify the Apraxia of speech. This work is carried out with the voice samples from the Ultra suit database of Apraxia of Speech. To highlight the advantages of the used NNs, the main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision The major characteristics of Apraxia of Speech include varying speech sound errors on consonants and vowels while repeating syllables and words, groping, or struggle, to have appropriate articulatory positioning (of lips, tongue, and jaw) for many of the sounds or words. So, speech processing plays a major role in identifying disfluency.

2. Proposed Algorithm:

Technology has the competence to increase motivation and engagement with therapy and to mitigate barriers associated with distance and access to speech pathology services, so ASRs place a major role in helping patients. Recognition of Apraxia of Speech is crucial since it exhibits several characteristic features. To create a better communication platform between the machine the process involves strenuous steps since the patients would take a quite long time to spell out the words and for the machine to get trained and tested. Figure 1 shows the flow of the work proposed to identify the Apraxia of speech.

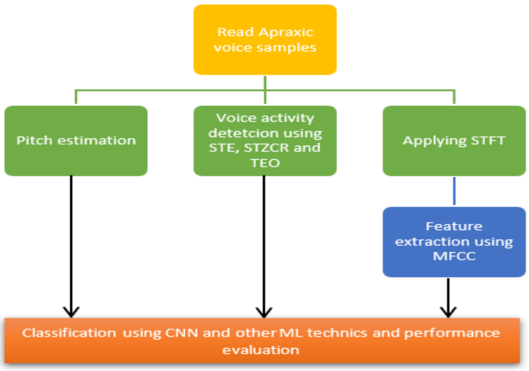
Read Apraxic voice samples
Pitch estimation
Voice activity detetcion using STE, STZCR and TEO
Applying STFT
Feature extraction using MFCC
Classification using CNN and other ML technics and performance evaluation

FIGURE 1: Proposed Methodology in the current work

The main contributions of this research are as follows:

- A system based on deep learning is proposed for the detection of AOS by using speech signals.
- It has been shown that a hybrid model consisting of Mel spectrogram, TEO, and CNN gives effective results in detecting AOS.
- By using Mel spectrogram, TEO with deep learning techniques, classification performance is increased.

The rest of this paper is structured as follows: In Section 3, the background information about the techniques used is described. The information about the experiment and results is summarized in Section 4, Section 5 gives the conclusion of the work, in section 6 the dataset used in the study is highlighted and in section 7 we have acknowledgment.

3. Background

3.1 Pitch estimation using autocorrelation

Pitch is one of the most important components in various speech processing systems. Pitch originates due to vocal fold vibration, and the frequency at which the vocal folds vibrate is the fundamental frequency. It is an important attribute of voiced speech. Periodicity related to voiced speech segments is determined as “pitch step” in the time domain and as “fundamental frequency” or F0 in the frequency domain. Speech classification into voiced and unvoiced (or silent) portions is important in many speech processing applications. In addition, segmentation of voiced speech into individual pitch epochs is necessary for several high-quality speech synthesis and coding techniques [5].

A frequently used method to evaluate pitch is grounded on detecting the highest value of the autocorrelation function in the region of interest. Given a discrete-time signal x(n), which is defined for all n, the auto-correlation function is generally given as [6]

$$R_x(m) = \lim_{n\to\infty}\left(\frac{1}{2N+1}\right)\sum_{n=-N}^{N} x(n)x(n+m) \qquad (1)$$

The autocorrelation function of a signal is a conversion of the signal that is useful for displaying structure in the waveform. Thus, for pitch detection of a nonstationary signal, such as speech, the long-time autocorrelation measurement will not be giving meaningful information. Thus, it is rational to express a short-time autocorrelation function, which operates on short segments of the signal as [7]

$$R_x(m) = \frac{1}{N}\sum^{N'}{}_{\substack{n=0\\0\le m\le M_0}} [x(n+I)w(n)][x(n+I+m)w(n+m)] \qquad (2)$$

here w(n) is the window for analysis, N is the section length which is analyzed, N’ is the number of signal samples utilized in the calculation of R(m), Mo is the number of autocorrelation points that should be computed, I represent the index of the frame. For pitch detection applications commonly set to the value in:

$N^{j} = N - m$
(3)

With the intention of having only the N samples in the analysis frame (i.e., x(I), x(I+1), . . .

, x(I+ N - 1)) are used in the autocorrelation mostly been used agreeing to a maximums (200 samples at a 10 kHz sampling rate)and a 30 ms analysis frame size. [7]

3.2 Voice activity detection (VAD)

Based on the value of estimated pitch, the End Point Detection of a dysfluency speech can be recognized effectively and it is contemporary speech information for further speech process- ing techniques. Enhanced performance and quality of speech coding, speech recognition, and speech synthesis are directly influenced by the success rate of endpoint detection. In recent works, it is evident that Short Term Energy (STE) and Short Term Zero Crossing (STZCR) are extensively used to perform EPD. In this work, we have made a comparative study on STE and STZCR along with Teager Energy Operator (TEO) methods to choose the best fit in the method for the current work.

3.2.1 Short Term Energy (STE)

Among several generic EPD algorithms, the STE and STZCR are best suited for the detection of dysfluencies in the time domain. These algorithms help in differentiating/ identifying voiced and nonvoiced regions in the given speech. The amplitude over unvoiced slices will be strikingly lower than that of the voiced slices. The amplitude distinction is mirrored by the short-time energy of speech signals. In a distinctive speech signal, it is observed that some of the properties noticeably change with time. For instance, we can perceive a significant distinction in the peak amplitude of the signal and a substantial variation of fundamental frequency within voiced regions in a speech signal. This evidence advocates that simple time domain analysis techniques are capable of providing beneficial information on signal features, such as intensity, excitation mode, pitch, and possibly even vocal tract parameters, such as formant frequencies. STE is the average of all the amplitudes of the speech signal. The STE at any given frame can be expressed as [8]

$$E = \sum_{n=-\infty}^{+\infty} s^2(n)$$
(4)

Where E represents the energy of the given speech signal, s(n) represents the discrete form of the speech signal. Between, voiced speech, unvoiced speech, and noise, the noise would have the lowest short-time energy, voiced speech would have maximum and the unvoiced speech would have intermediate.

3.2.2 Short-Term Zero Crossing (STZCR)

Grounded on this fact, It would be easy to distinguish the voice segments and the noise segments. Further down at high SNR conditions, short-time ZCR can be employed to distinguish the unvoiced and voiced speeches. The STZCR signifies the times a speech frame crosses the horizontal axis a greater number of crossings signifies a non-voiced region while fewer zero-crossings signify a voiced region [1] in the current work-frame size is considered to be 10ms to 30 ms and ZCR is computed with half-frame size.

The figure 2 shows the example of speech segmentation and the respective sequence of STZCR [9].

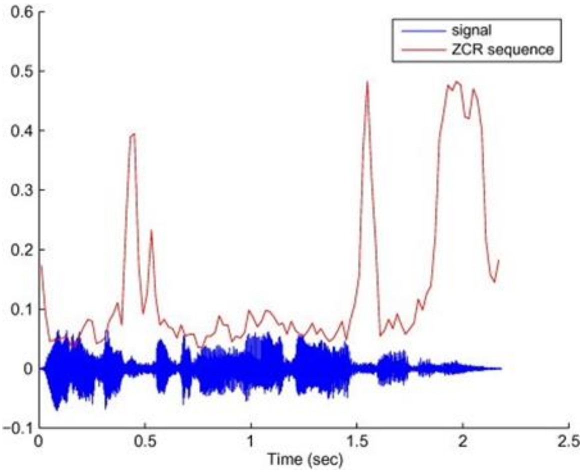

FIGURE 2: Example of speech segmentation and the respective sequence of STZCR

The STZCR can be well-defined bestowing the following expression

$$Z_{(i)} = \frac{1}{2X_L} \sum_{n=1}^{X_L} |sgn[x_i(n)] - sgn[x_i(n-1)]|$$
(5)

Where Sgn(.) is a Sign function expressed as

$$sgn[x_i(n)] = \begin{cases} 1 & x_i(n) \geq 0, \\ -1 & x_i(n) \geq 0 \end{cases}$$

Algorithm 1

Algorithm for classification of Voice and unvoiced region of speech

1. Input the speech signal
2. Perform framing and signal processing
3. Compute the short-term zero crossing rate
4. If the STZCR is a smaller value then it’s to be identified as a Voiced region else Unvoiced region

3.2.3 Teager Energy Operator (TEO)

The STE and STZCR fail in extracting the values of contours in real-time, hence TEO is more suitable for improving the efficiency. The pivotal part of the speech recognition system, when the speech is observed in a noisy environment, is speech endpoint detection. The recognition performance also computational complexity of the speech recognition system depends on the accuracy of detection. The work utilizes an endpoint detection of speech signals based on the Teager Energy Operator (TEO). To safeguard the accuracy in a noisy environment and also the sturdiness to changes in absolute levels this method makes use of three state transition and judgment mechanism which is established on double thresholds.

The characteristics of TEO: For a continuous-time signal s (t), TEO is defined as:

$$\psi\ (s\ (t)] = (s^j(t))^2 - s^j\ (t)\ s^{jj}(t) \qquad (6)$$

Which, $s^j(t)$= ds(t) / dt , s (t) is the continuous time domain signal. To describe a discrete-time signal, equation (2) can be approximated as:

$$\psi[s(n)] = s(n)^2 - s(n-1)s(n+1) \qquad (7)$$

Where: ψ [s (n)] represents signal TEO, s (n) is the identified to be a sample of discrete signal time in n.

$$\psi\ [s(n)] = s(n)^2 - s(n-1)s(n+1)=A^2\Omega^2 \qquad (8)$$

It can be seen from the above equation, that the output of TEO not only includes both amplitude information (A) but also contain frequency-domain information **Ω**.

The noise characteristics of TEO: As a next step, the noisy signals are converted into frames, and windowing is performed on each frame as per the below equation (5) The TEO of the clean speech signal s(n) can be represented as x(n) :

$$\psi\ [x(n)]= \psi\ [s(n)]+ \psi\ [\omega\ (n)]+2\ \psi\ [s(n),\omega(n)] \qquad (9)$$

Smoothing: as mentioned earlier TEO is mainly advantageous for speech recoded in noise. So, to reduce false-positive values which will be the effect of jitter noise we make use of smoothing in turn robustness of the detection will be improved. Below equation [6] detects a mathematical representation of the smoothing process involved.

$$E[\psi(n)]' = \frac{E[\psi(n-1)]+E[\psi(n)]+E[\psi(n+1)]}{3} \qquad (10)$$

As stated at the beginning of this section comparison of TEO and STE in the detection of the Endpoint is shown in below figure 3.8. and TEO can perform better in the detection of endpoints.

3.3 Feature Extraction

Feature extraction is an admissible process in the achievement of the ASR system. A pleasing technique will be capable of characterizing vital attributes and also confiscating irrelevant characteristics. In ASR systems acceptable classification is derived from excellent and quality features. A few of the speech feature extraction techniques are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP). This work exploits the MFCC-based feature extraction technic. The utmost widespread and foremost method used to extract spectral features is calculating Mel- Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques applied in speech recognition based on frequency domain exploiting the Mel scale which is built on the human ear scale. Compared to that time-domain features MFCCs are measured as frequency domain features which are very precise. [9], [10]. Studies have proved that human insight into the contents of the frequency of sounds for speech signals will not be a linear scale. Therefore, for an individual tone with a real frequency, f, measured in Hertz, an independent tone is calculated on a scale known as the “Mel scale. The Mel frequency scale is nothing but the spacing of linear frequency below 1000

Hertz and a logarithmic rate above 1000 Hertz. As a point of locus, the pitch of a tone of 1 kHz, 40 dB above the threshold of auditory perception, is distinct as 1000 Mels. Therefore, the following estimated formula can be used to determine the Mels for a given frequency f in Hz [4]:

$$m = 2595 \log_{10}(\frac{f}{100} + 1) \qquad (11)$$

where, f = frequency in Hz; dB = decibel; Log= logarithm factor [4].

Such a measurement is used because it is not to be mistaken for the human ear’s familiarization with a sound. In the conventional feature extraction technique, MFCCs can be calculated as the result of successive weighted summation of all the 8 Mel-frequency coefficients obtained from 8 frequency bands of 20 Hz at each half band.

In the proposed work 20 Mel-frequency coefficients were obtained from 20 frequency bands at 20KHz of the sampling rate. To obtain the transformation of signal from the time domain to the frequency domain, Discrete Fourier transform (DFT) has been applied the mathematical relationship to describe this transformation

As a part of next stage, there arises a need of finding the power of the spectrum. The power spectrum is obtained as

$$P_i(k) = \frac{1}{N} |F_i(k)|^2 \qquad (13)$$

Further, the power spectrum P(k) is passed through a series of Mel triangular filer windows to find the MFCC spectrum. The triangular filter frequency is calculated by

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le kf(m+1) \\ 0 & k > f(m+1) \end{cases} \qquad (14)$$

The logarithm energy spectrum of each frame having f(m) as the center frequency of the mel triangle filter is obtained using a logarithmic procedure

The $S(m) = \ln[\sum_{k=0}^{N-1} P(k) H_m(k)], \qquad 0 \le m \le M$ (15)

P_m represents the power spectrum, H_m(k) describes the filter window and M is the number of filter windows. Below figure is the depiction of speech sound signal and its MFCC spectrum.

3.4 Classification of apraxic and non-apraxic speech

Convolutional Neural Network Technically, the CNN model comprises the following layers pooling layer, convolution layer, fully connected layers, and classification layer [11]. The below figure 3 shows the architecture of a CNN.

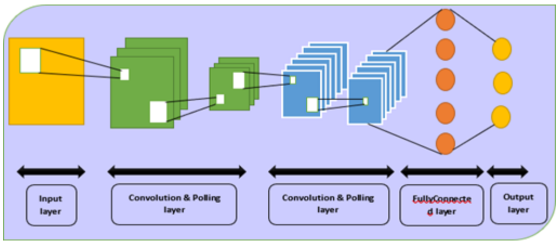

FIGURE 3 Basic architecture of CNN

The convolution is the first layer from which an input image derives features. The convolution is a filter applied to the input image to extract a feature map from the input image. The height and weights of the filters are smaller than the input volume. The formula for the convolution process is assumed in Equation 1. The input image and kernel are denoted by f and h, respectively. The row and column indexes of the result matrix are represented by m and n, respectively.

$$G[m,n] = (f * h)[m,n] - \sum_j \sum_k h[j,k] f[m-j, n-k] \qquad (16)$$

The pooling layer is another architecture of pre-trained deep networks. The pooling layer is used to reduce the dimensions of the input image after the convolutional layer and to accelerate the calculations. Fully connected layers are an important constituent of deep networks that have proven to be very successful in identifying and classifying images. The completely linked input layer takes and flattens the output of the previous layers after turning them into a single vector that can be an input for the next level. The last layer is the output layer and probabilities are estimated for each label. Softmax is generally selected in this layer. The softmax formula is computed in Equation 17.

$$Softmax(x)_j = \frac{e^{xi}}{\sum_{n-1}^{N} e^{x_n}} \quad n = 1, j = 1 \ldots .N \qquad (17)$$

The evaluation criteria outlined below were used to calculate the results received from the suggested technique. Accuracy is measured by the number of accurate predictions, divided by the total number of predictions. The accuracy evaluation is given in Equation (18).

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|} \qquad (18)$$

Precision is an evaluation that predicts the probability that a positive forecast is correct. The precision evaluation is given in Equation (19).

$$Precision = \frac{|TP|}{|TP| + |FP|} \qquad (19)$$

Sensitivity is the true positive of the predicted data belonging to the positive class. Sensitivity is calculated as shown in Equation (20)

$$Sensivity = \frac{|TP|}{\mathrm{TP}|+|\mathrm{FN}|} \qquad (20)$$

Training of CNN was carried out using the Adam Optimizer [Kingma and Ba, 2014], 32 is the size of minibatch and the number of epochs was 50. The proposed work was trained on 80% of each dataset and tested against the remaining 20% of that dataset.

4.1 Experiments and Result

The proposed work was tested on UltraSuite Repository where we have 4 speakers with apraxia of speech below are the details of the speakers

TABLE 2 Data set details of Apraxia of speech

SL. No.	GENDER	AGE-Y	AGE-M	AGE	SSD SUBTYPE
1	Female	10	11	10.92	childhood apraxia of speech
2	Male	8	11	8.92	childhood apraxia of speech
3	Male	10	2	10.17	childhood apraxia of speech
4	Male	13	4	13.33	childhood apraxia of speech

The outcomes of the proposed work are explained in this section. As per the literature survey conducted this word is itself one of its novel kind since there are no literatures found to be presenting AI techniques for the detection and classification of the disfluency Apraxia of speech. Throughout this paper the apraxic and normal pronunciation of word “data” is being referred to report the exhibits of the research work. The speech signal is read and at first, the peaks of the speech signal are extracted in time domain and it is evident that the apraxic voices exhibit more peaks and distribution is not as per the magnitude of the speech signal as observed from figure 4 and figure 6.

In the figure 6 it can be observed that the peaks are distributed as per the variation in the amplitude of the voice sample, whereas the apraxic sample exhibits the too many peaks even

TABLE 3 Tabulation of number of peaks detected for normal and apraxic speeches.

Word	Normal	Apraxic
Celebration	84,58,482	1,17,77,156
Data	74,91,628	1,40,37,192
Debt	27,60,145	1,46,82,553
Invitation	1,48,29,518	3,38,65,110
Shine	31,36,709	43,72,995

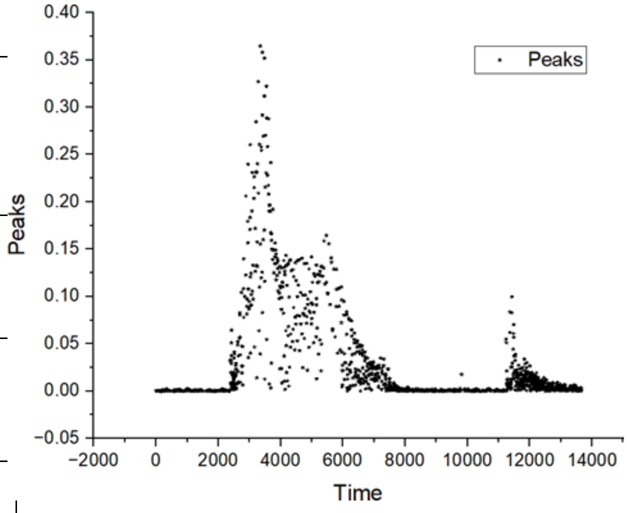

FIGURE 4 Corresponding peaks detected Apraxic speech

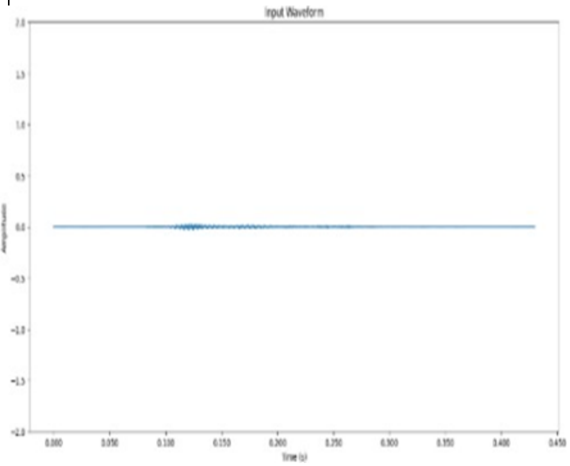

FIGURE 5 Original Apraxic sample

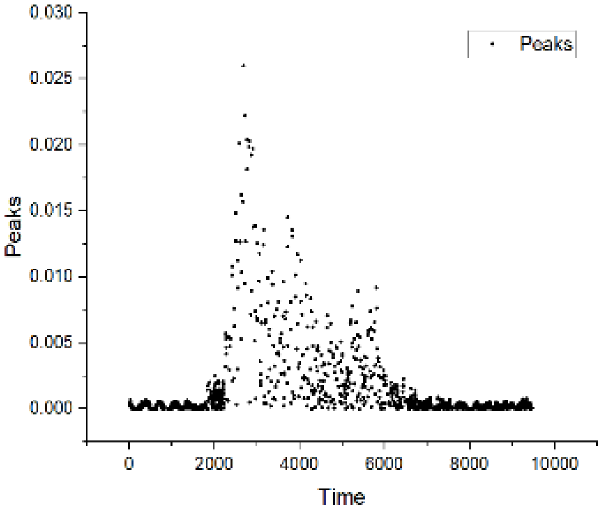

FIGURE 6 Corresponding peaks detected Normal speech

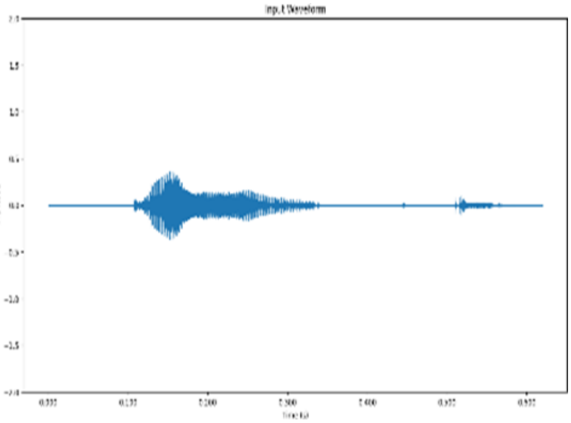

FIGURE 7 Original normal speech

when the amplitude variations in the audio signal is less, which is observed in figure 5. This abnormality in peaks are help use in identification of apraxic and healthy voice samples.

To perform further study we have extracted some speech features from samples, namely 20 MFCC features, pitch, roll-off, STZCR, TEO, and spectral centroid. To extract MFCC the signal of converted from the time domain to the frequency domain Using STFT.

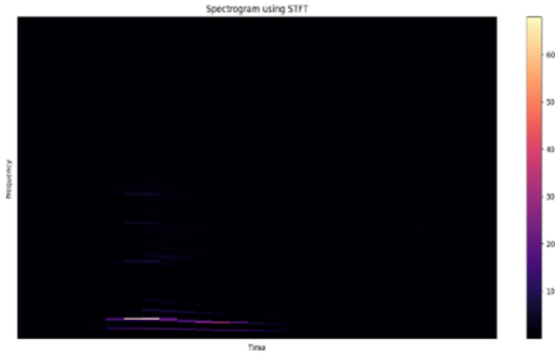

FIGURE 8 STFT spectrum of Apraxic speech

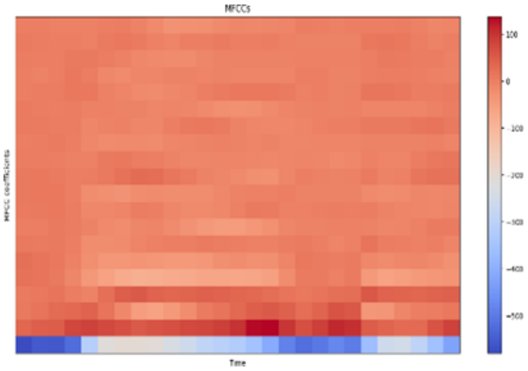

FIGURE 9 MFCC co efficients of Apraxic

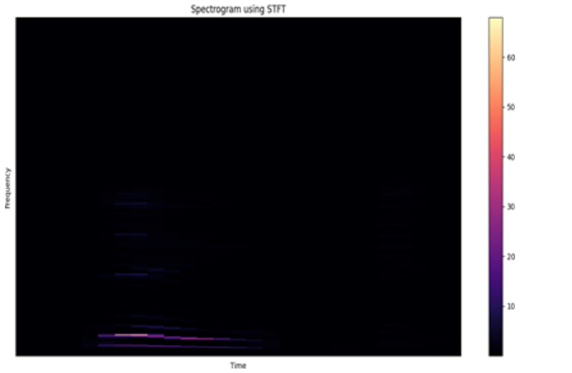

FIGURE 10 STFT spectrum of Normal speech

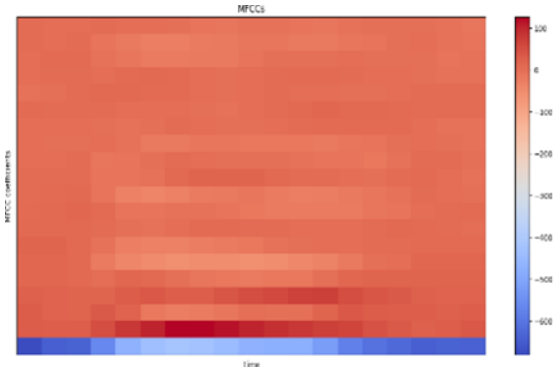

FIGURE 11 MFCC co efficients of Normal

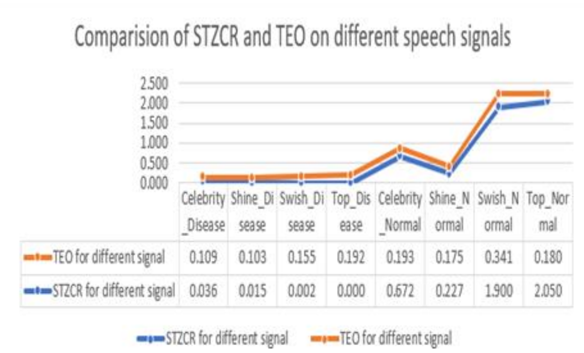

CHART 1 Comparison of STZCR and TEO for different speech

The sequence of Fourier transforms of a signal being windowed will form a new transformation technic popularly known as the short-time Fourier transform (STFT). Typically, STFT delivers the time-localized frequency information for conditions in which frequency components of a signal diverge over time, although the typical Fourier transform provides the frequency information averaged over the total signal time interval. The below figure shows the STFT of the 2 input signals healthy voice and apraxic voice respectively.

Further, the extracted feature STFT, RMSE, roll-off, spectral centroid, MFCC, TEO, and STZCR are fed to different classifiers. In the proposed work we have carried out a comparative study on the efficiency of different classifiers. The classifiers that are under consideration are some of the machine learning and deep neural networks. The ML technics that are considered are linear SVM, Radial SVM, Logistic regression and KNN mean while the DNN used is CNN. The comparative analysis of different Ml algorithms while considering STZCR and TEO exhibited in the figure 12.

FIGURE 12 Comparison of different ML algorithm-based classifiers considering different VAD methods

The important observation, while the classification was performed, is the VAD techniques considered makes an influence on the efficiency of the classifier. The best results are obtained when the STZCR is being considered rather than the TEO. The obtained results are shown in the Figure 13 along with the sample confusion matrix generated for the same. The number of epochs is fixed to be 150 with a batch size of 64 based on our trial-and-error process to get the maximum efficiency. Below table shows the progression of recall, precision and specificity of the model for Apraxic word Disease. Over all we are able to achieve average of 87.5% of efficiency in testing phase when STZCR of apraxic word is considered and when TEO of the apraxic word is considered the test efficiency is found to be 78%. Similarly, for normal speech considering STZCR and TEO we are able to achieve 78% of efficiency. Figure 7. Shows the results obtained for performance evaluation CNN

By comparing the efficiency from figure 6 and figure 7 it is evident that the DNN method is suitable for identification and classification of Apraxia of speech. Whereas the ML methods are poorly performing for the classification of the apraxic disfluency.

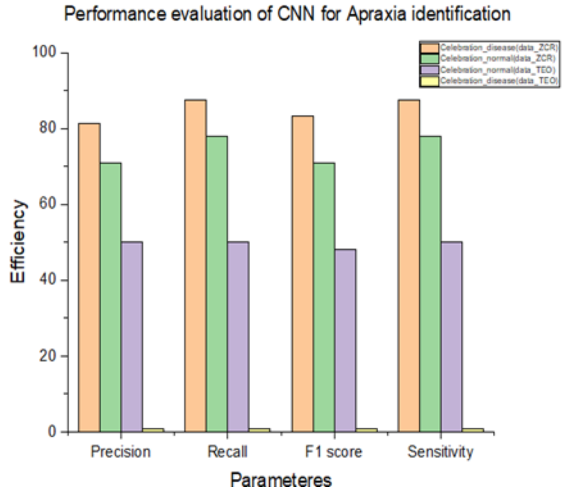

FIGURE 13 Performance of CNN considering different VAD methods

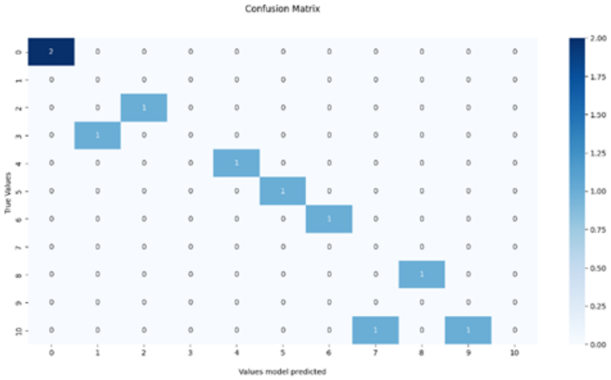

FIGURE 14 Sample Confusion matrix generated through CNN for a apraxic word classification

5.1 Conclusion

Identification of Apraxia of speech in the early stage plays a crucial role in helping the affected patients to a greater extent. In this work, we have presented a comparative analy- sis of ML and DL models' performance in detecting Apraxia of speech. Through the proposed work inference can be drawn that the STZCR voice activity detection method combined with the Deep neural network like CNN can give better accuracy in the identification of the Apraxia of speech. Further, we would like to extend our work to increase efficiency to a greater extent.

Data Availability

The database used for this particular study is open source data available at this link Ultra- Suite Repository (16)

Acknowledgment

We acknowledge the support from REVA University for the facilities provided to carry out this research work.

References

[1] K Knollman-Porter. Acquired apraxia of speech: a review. *Top Stroke Rehabil*, 15(5):484– 93, 2008.

[2] J Ogar, H Slama, N Dronkers, S Amici, and M L Gorno-Tempini. Apraxia of speech: an overview. *Neurocase*, 11(6):427–459, 2005.

[3] R T Wertz, L L Lapointe, and J C Rosenbek, Apraxia of speech in adults : the disorder and its management, 1984.

[4] Apraxia of Speech and a Case Example - Better Speech, 2022. https://blog.betterspeech.com/post/apraxia-or-apraxia-of-speech-and-a-case-example.

[5] S, Ajibola Alim, and N. Khair Alang Rashid. Some Commonly Used Speech Feature Extraction Algorithms. *From Natural to Artificial*

Intelligence - Algorithms and Applications, 2018.

[6] DNN-based Causal Voice Activity Detector, 2022. https://www.researchgate.net/ publication/315955578_DNN-based_Causal_Voice_Activity_Detector

[7] Pitch detection algorithm: autocorrelation method and AMDF, 2022. https://www. researchgate.net/publication/228854783_Pitch_detection_algorithm_autocorrelation_ method_and_AMDF

[8] D S Shete and S B Patil. Zero crossing rate and Energy of the Speech Signal of Devanagari Script. *IOSR Journal of VLSI and Signal Processing (IOSR-JVSP*, 4(1), 2014.

[9] Zero Crossing Rate - an overview | ScienceDirect Topics, 2022. https://www.sciencedirect. com/topics/engineering/zero-crossing-rate

[10] M B Er, E Isik, and I Isik. Parkinson's detection based on combined CNN and LSTM using enhanced speech signals with Variational mode decomposition. *Biomedical Signal Process- ing and Control*, 70:103006–103006, 2021.

[11] D C C Cire¸san, U Meier, J Masci, L M Gambardella, and J Schmidhuber, High-Performance Neural Networks for Visual Object Classification, 2011.

[12] B Jan. Deep learning in big data Analytics: A comparative study. *Computers and Electrical Engineering*, 75:275–287, 2019.

[13] C Chen, Z Hua, R Zhang, G Liu, and W Wen. Automated arrhythmia classification based on a combination network of CNN and LSTM. *Biomedical Signal Processing and Control*, 57, 2020.

[14] M A Little, P E Mcsharry, E J Hunter, J Spielman, and L O Ramig, Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," Nature Precedings, 2008.

[15] I Bhattacharya and M P S Bhatia. SVM classification to distinguish Parkinson disease patients. *Proceedings of the 1st Amrita ACM-W Celebration of Women in Computing in India, A2CWiC'10*, 2010.

[16] A Eshky, UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions.