ZKG INTERNATIONAL

ISSN: 2366-1313

Modern deep learning algorithms for speech recognition algorithm reliability evaluation

Garaga Srilakshmi[1], Dr. Alok Agarwal[2], Dr.Dola Sanjay S[3]

Research scholar ECE, JJTU, Rajasthan.

Guide, Department of Electronics and Communication Engineering, JJTU, Rajasthan.

Co-Guide, Principal Aditya College of Engineering And Technology, Surampalem. A.P

ABSTRACT

The quality of a user's recorded voice may be improved with the use of a speech enhancer programmer. The input voice may be isolated from background noise thanks to the speech augmentation technique. Research into voice enhancement aims to uncover methods for rescuing speech that has been corrupted by background noise. This breakthrough has a wide range of potential applications, including but not limited to teleconferencing, speech recognition, mobile phones, and hearing aids. The potential for speech enhancement technology has grown significantly with the introduction of smart phones, smart watches, and other wearable and augmented reality devices. The intelligibility of remote command and control might be improved with the use of speech augmentation technologies. The purpose of today's voice-enhanced smart devices is to learn more about their users. In the future, it may be possible to improve the amplification quality of voice-activated devices. However, current speech enhancers can only eliminate quasi-stationary noise because of computational limitations. After the noise has been removed, the remaining signal's strength may start to degrade.

If language identification and customization techniques were made accessible, more people could begin wearing hearing aids. It is possible to use the phrase "speech enhancement" to refer to any method used to improve the quality of transmitted sound. One goal of these methods is to reduce listener fatigue and increase speech intelligibility. In this case, it might be helpful to utilize synthetic voice to drown out any ambient noise.

I. INTRODUCTION

Speech refers to any kind of communication in which only verbal sounds are exchanged. The meanings of words and voice cues change based on context, the speaker's language, and the listener's mood. All of the aforementioned elements of verbal communication are used on a daily basis in order to accomplish a wide range of goals. Programming a computer to recognize and respond appropriately to human speech is challenging. Despite its numerous advantages, the speech or verbal communication processing arrangement system struggles to function in a noisy environment. There are a variety of obstacles that prevent robots from understanding human speech at the present time. Existing speech-based or speech-arrangement systems, however, have difficulty functioning in noisy or emotionally charged environments. A computer has to understand Assamese speech just like a person does, complete with the appropriate emotional feelings, before even the most advanced verbal communication speech or verbal communication interface can be developed. Everyone in this group has above-average listening and response skills. In Assamese, one's vocal tone may be indicative of one's character and personality. Developing and implementing a plan for deploying some kind of Integration of message, speaker characteristics, linguistic identity, and intended affective attitudes is essential in crafting an effective speech or verbal communication tool. Those who depend heavily on computers and other electronic gadgets would be crippled without it. Accurate Assamese voice translation of written conversations requires advanced capabilities like emotion detection and identification. This is the proper way to begin a joke or explain your motivation. Phoneticians analyze speech patterns to decipher what's being said and how someone feels. In order to better comprehend and organize systems that can deal with Assamese emotional speech feelings in connection to the conveyed message, the creation of an Assamese emotional speech tool is necessary [2]. For (a) Assamese speech or speech-like communication to be analyzed and (b) emotional feelings to be recognized, a system must be built. And (b) making full use of existing tools to create an Assamese-language speech or speech-like communication synthesizer with the ability to generate the appropriate emotional prototype experiences. Therefore, it may be possible to classify or identify the emotional speech feelings needed to comprehend Assamese verbal communication speech. A synthesis of text including emotional experiences, or an emotional synthesizer, might be developed for use in Assamese verbal communication speech, given the information provided in all meanings. Training for identifying these emotions in people's voices may benefit from the models used by emotion recognition software. It has been trained to understand nuances and subtleties in human speech. The purpose of this research is to isolate and classify the many ways in which human voices convey affect in recorded speech. It seems to be difficult to put one's feelings into words. Writing in this style incorporates the author's name, the message's context, and the author's native tongue. The regularity with which native Assamese speakers use metaphor depends on a number of contextual circumstances.

II REVIEW OF LITERATURE

Feature extraction is a critical processing step required by all speech recognition systems. At this step, the fundamental speaker characteristics have been retrieved from the raw sound stream. It is common practice for feature vectors to include both speaker and audio data. Speech recognition systems depend exclusively on input from the speaker during training. After the speech signal has been preprocessed, numerous metrics regarding the speakers may be recovered. Mel frequency cepstral coefficients (MFCC), formants, and the Normalized Pitch Frequency (NPF) coefficient are all instances.

Linear Prediction Cepstral Coefficients (LPCCs) are a popular feature parameter. The Minimum Feature

Coherence (MFCC) is another widely used metric. LPCC features have been there since the '60s (Atal et al. 1978), but their broad use is a result of improvements in efficiency, speed, and hardware implementation. Reliability, classification accuracy, and other attribute all declines under noisy situations, rendering LPCC ineffective in practice.

Since quite some time (Wang & Lawlor 2017; Milner & Darch 2011), MFCCs have been employed in speech and speaker ID systems. All Cepstrals have the same set of MFCCs, which are a collection of minor but identifying features. The MFCC also has the additional advantages of being steady and unaffected by ambient noise (XinXing & Xu, 2012). Polar and Miller (2005) utilized HMM to construct features from the Fast Fourier Transform (FFT), Linear Prediction Coefficient (LPC), and Mel-Frequency Characteristics (MFCC) for the purpose of speech recognition. The modified Mel filter bank proposed by Kopparapu and Bhuvanagiri (2013) depends on the connection between the MFCC characteristics of raw sampled speech and re-sampled speech. Using algebraic parametric structures learnt on raw audio, Valentini-Botinhao et al. (2014) suggested a technique for adapting MFCC.

Although great progress has been achieved in the area of speaker recognition, various flaws remain. It's likely that developments in technology may increase the SR method's identification accuracy.

To attain their aim of text-free speaker identification, Ma et al. (2016) employed the Histogram Transform (HT) feature in combination with MFCC characteristics. This implies that alternative state-of-the-art models may be quickly integrated into HT for SI. HT employs random transformations to offer extra training data for histogram PDF estimation, therefore eliminating the discontinuity issue inherent in the production of multiclass histograms. This HT technique works better than GMM at detecting speakers, but it becomes prohibitively costly to operate as the number of speakers rises.

Kim et al. (2016) presented the feature extraction approach known as Power Normalized Cepstral Coefficients (PNCCs). Learning to recognize key elements is a fundamental issue in pattern recognition. The temporal and frequency properties of the audio stream, as well as the available data storage space, determine the selection of which feature to utilize in an SR system. What drives the PNCCs is the power-law nonlinearity that is embedded into the ear itself.

Using neural networks and a wrapper method, Mongrel Kabir et al. (2010) established a novel strategy for picking features. This strategy may be used to reduce down an SR system's feature set without significantly harming its efficiency. The suggested feature set could be enhanced by employing more granular wrappers. Improving the SR scheme's efficiency is a difficult process.

Speaker verification in the research of Vasamsetti Srinivas and Santhirani (2019) was performed using a cosine distance calculation-based scoring technique, while speaker recognition (SR) was accomplished using an AFB-based SVNN. The SI response of the SVNN scheme is enhanced with the application of spectral features (MFCC, Kurtosis, Skewness, and Autocorrelation) and the AFB approach. Raising the detection rate involves tweaking the NN's weight.

Using short spectrograms as input to CNN to test its speech-recognition effectiveness was suggested by Lukic et al. (2016). The BNF and the front-end standalone SE are both trained using a Deep Learning model (CNN). In CNN's deeper layers, you may train a number of classifier back ends. This approach enables an SR system to be trained from scratch. Because of this strategy, less labor is required to extract and pick features in ASI. When there is a big quantity of training data to process, system operation becomes more time-consuming and labor-intensive.

III SYSTEM ANALYSIS

Depicts the distributions of the input voice signals after the gaps have been removed. Typical and very lengthy input sounds, as well as running speech, are used to evaluate the muting algorithm's efficacy. The signal-to-noise ratio (SNR) has varied effects on the median output when dealing with healthy and abnormal samples of input voices. The quality of the input speech samples also had an effect on the SNR figures.

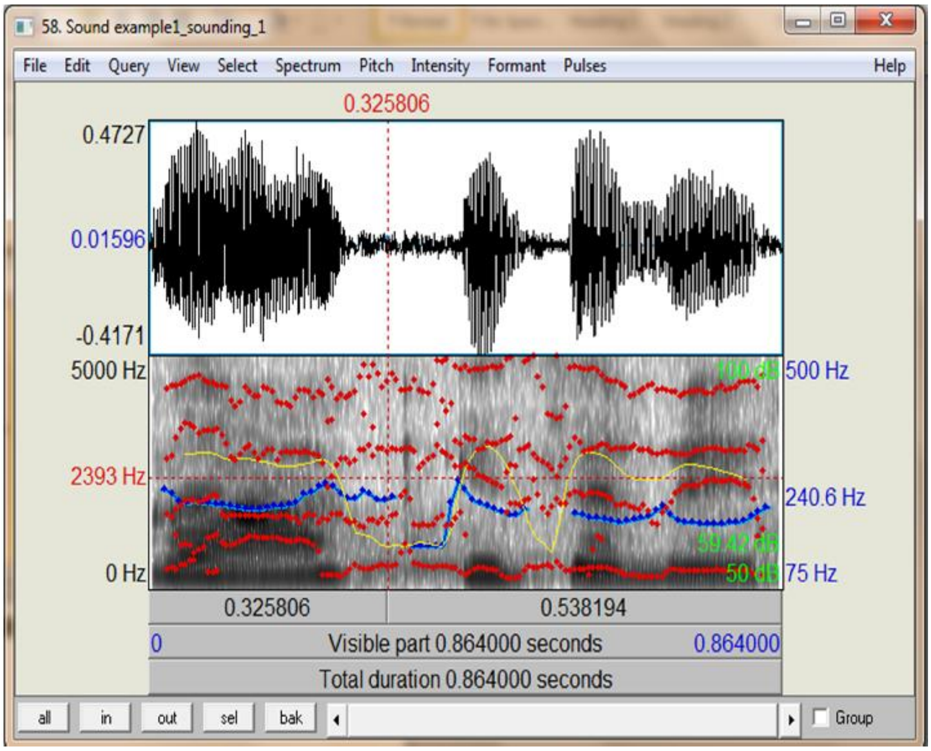
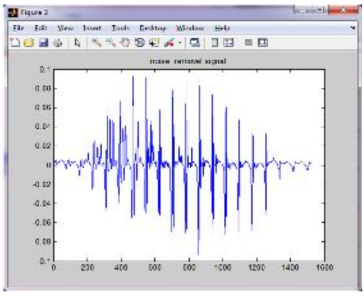


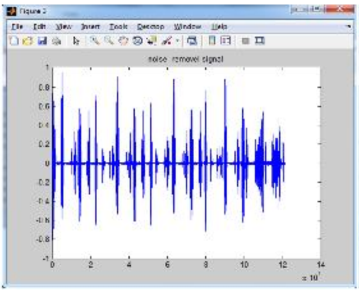
Figure 4.10. Noise Waveform for Input Voice Signals

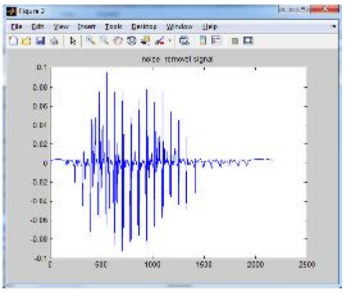
Figure 4.10 depicts the distribution of the input noise waveform for normal and pathological preprocessing stages, which are analogous to the silence removal operation.

The input speech signal of the nine running voice samples (1-9) was used to compare healthy and ill voices, and noise reduction was more effective than silence removal (Figure 4.12). When noise was reduced, the median signal-to-noise ratio (SNR) was better than when silence was raised.

Given the striking dissimilarities between the speech samples used for noise cancellation and those utilized for noise retention, these results were to be anticipated.



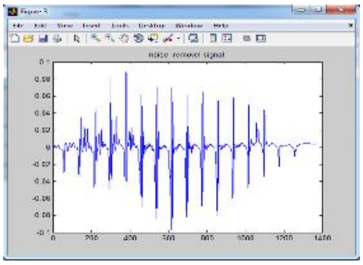




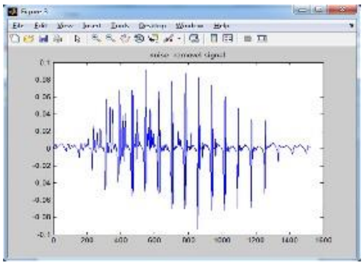
(a) **Noise subtractionSignal Sample 1**

(b) **Noise taking awaySignal Sample 2**

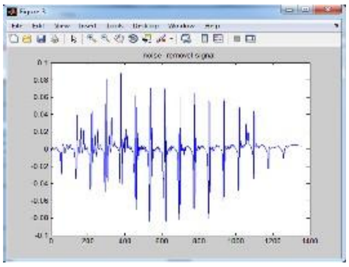
(c) Noise RemovalSignal Sample 3



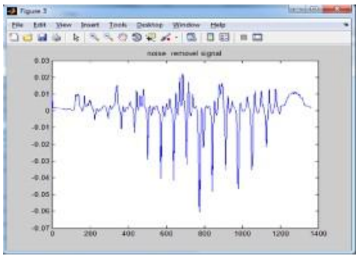
(d) Noise RemovalSignal Sample 4



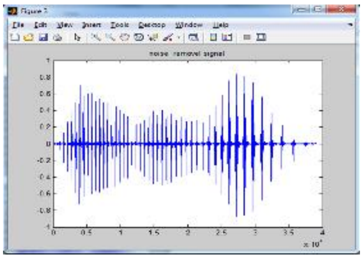
(e) Noise RemovalSignal Sample 5



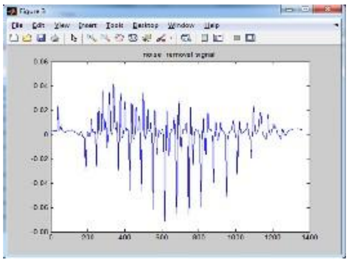
(f) Noise RemovalSignal Sample 6



(g) Noise RemovalSignal Sample 7



(h) Noise RemovalSignal Sample 8



(i) Noise RemovalSignal Sample 9

Figure 4.11 (a) to (i) Noise Removal for Input Voice Signals

Figure 4.11 depicts examples of continuous speech with and without noise reduction applied, as well as voice signals that are both healthy and disturbed. The parameter of the recovered signal was lowered when applied to speech samples, but the resulting voice was superior to the original. The signal-to-noise ratio (SNR) of each sample is calculated after the input voice signal has been cleaned up.

Signal-to-noise ratios (SNRs) for the clean and unclean signals utilized in the SVD dataset are shown in Figure 4.13. Results show that SNR ratio is useful.

A. SNR Values for SVD Dataset

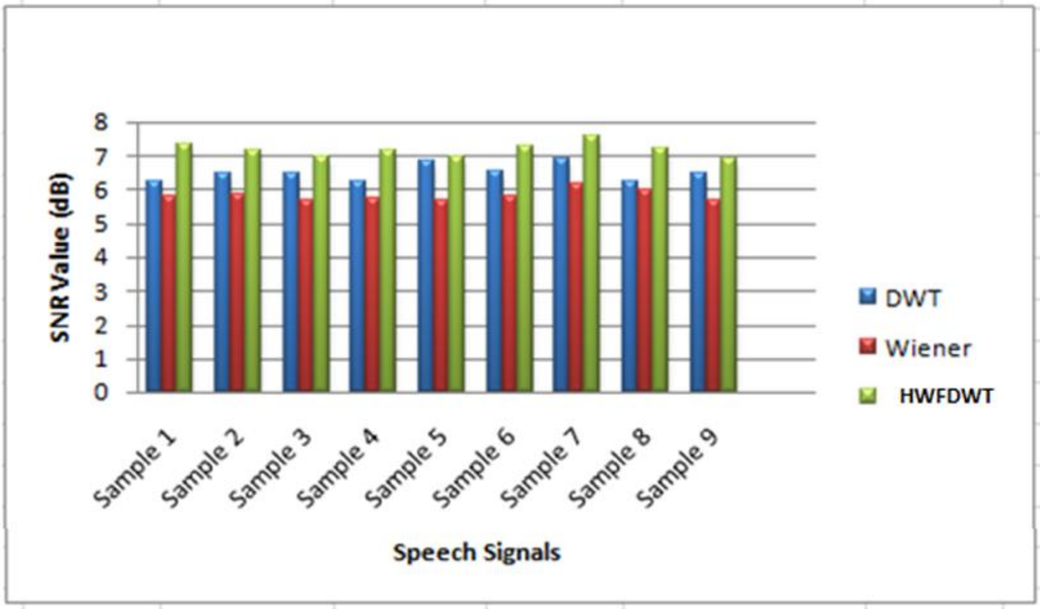


Figure 4.12. SNR Values of SVD Dataset

Figure 4.12 displays the SNR values extracted from the SVD dataset, broken down into healthy and abnormal speech signal distributions following the removal of background noise and quiet. On a sample-by-sample basis, the signal-to-noise ratio is calculated. Hybrid Wiener Filter Discrete Wavelet Transforms (HWFDWT) benefit from the SVD Dataset's preprocessing function, which involves comparing sample values with the Wiener filter, DWTs, and HWFDWT. The signal-to-noise ratio (SNR) of typical and disordered human speech varies.

B. SNR value for Private Real-time dataset

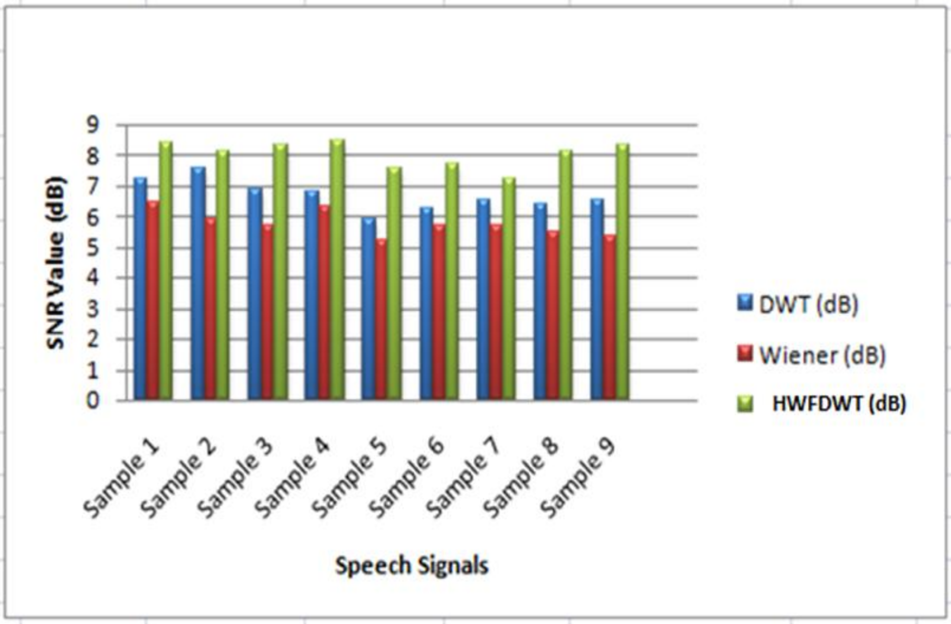


Figure 4.13 SNR Values of Real-Time Dataset

From the private real-time dataset, Figure 4.14 displays the SNR values and distribution of the muted/reduced signal. Each sample is evaluated independently to establish the signal-to-noise ratio. Several discrete wavelet transforms, the Wiener filter, and a hybrid of the two called the Hybrid Wiener Filter Discrete Wavelet Transform (HWFDWT) are applied to the sample values and compared to one another. Thanks to HWFDWT's preprocessing, the private real-time dataset has a high SNR across the board for all samples. Human speech, both normal and disturbed, has varying signal-to-noise ratios (SNRs).

IV SPEAKER IDENTIFICATION AND VERIFICATION

Table 5.11 Performance analysis of different classifiers for experimental dataset

Classifiers	Evaluation Measures				
	Accuracy	FA	FR	Specificity	RMSE
SVM	0.8753	0.030	0.4846	0.9042	1.0153
RF	0.9040	0.021	0.3873	0.9372	0.9100
NN	0.8924	0.026	0.4575	0.9147	0.9424

NN-SVM	0.9116	0.019	0.0126	0.9469	0.5959
DNN	0.9510	0.018	0.0128	0.9656	0.3536
RF-SVM	0.9720	0.012	0.0105	0.9968	0.1940

Combining RF with SVM results in a classifier with a remarkable 1% error rate, 97% accuracy, and 96% specificity for the observed data. Tables 3.7, 3.8, 3.9, and 3.11 provide the classification outcomes for the various datasets. On all four datasets, the RF-SVM achieves better performance than baseline methods, with FARs of 0.0018, 0.0016, 0.0019, and 0.012, and FRRs of 0.019, 0.027, 0.018, and 0.0105, respectively.

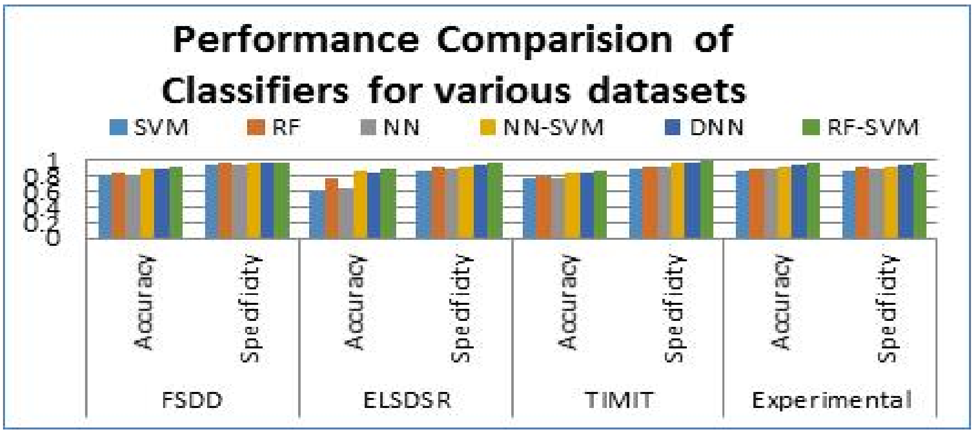
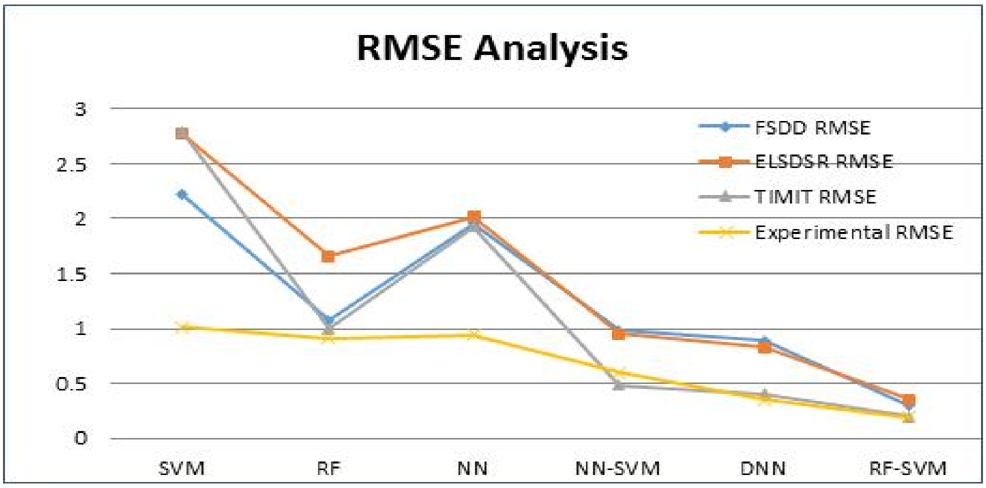


Figure5.10 Evaluation of the classifier's sensitivity and specificity over a range of datasets

Performance of ML classifiers on numerous datasets with varying degrees of imbalance is shown in Figure 3.15. The average accuracy and specificity of the recommended RF-SVM classifier are 91.65% and 98.38%, respectively.

Figure 3.16 displays the classifiers' recognition error rates as the root mean square error. Figure 3.16 shows that the spectral feature set coupled with the RF-SVM outperforms the separate statistical classifiers when using a minimal error function for recognition.

Statistics using F1_Score and Cohen's kappa



Cohen's kappa is used to evaluate performance in multi-class identification by considering both the observed and expected accuracy of the model (Ben-David, 2007). This attempts to account for the inherent bias in assessment by taking into account the fact that voiceprints are not distributed uniformly. Increased classification precision may be achieved by lowering the False Negative rate and increasing the True Positive rate. To reduce the impact of social class differences, Cohen's kappa and the F1 Score may be the most popular measures. Cohen's kappa and f1_score may be utilized to compensate for the assessment bias caused by the speaker data asymmetry by concentrating on the model's anticipated accuracy. The f1_score may be thought of as a sum of the scores for accuracy and recall.

Table5.12 The f1_score and kappa analysis of various classifiers

Classifier	Dataset(DS)	Kappa	f1_score
SVM	Experimental DS	0.8276	0.7972
	FSDDDS	0.8121	0.7158
	ELSDSRDS	0.7415	0.7536
	TIMITDS	0.7995	0.6652
RF	Experimental DS	0.9356	0.8857
	FSDDDS	0.9121	0.9120
	ELSDSRDS	0.8567	0.8432
	TIMITDS	0.9005	0.8327
	Experimental DS	0.7789	0.8732
	FSDDDS	0.8450	0.7945

ISSN: 2366-1313

NN	ELSDSRDS	0.7665	0.7917
	TIMITDS	0.8157	0.7427
NN-SVM	Experimental DS	0.9174	0.9032
	FSDDDS	0.9151	0.8917
	ELSDSRDS	0.8707	0.8538
	TIMITDS	0.8821	0.9017
DNN	Experimental DS	0.9207	0.9107
	FSDDDS	0.9078	0.9012
	ELSDSRDS	0.8816	0.8895
	TIMITDS	0.9076	0.9142
RF-SVM	Experimental DS	0.9314	0.9395
	FSDDDS	0.9187	0.9251
	ELSDSRDS	0.8992	0.9119
	TIMITDS	0.9104	0.9235

Table 5.12 highlights the performance of the RFSVM classifier against various existing methodologies in terms of Cohen's kappa evaluation metric and the f1_score. The proposed hybrid ML classifier produces higher inter-rater reliability of 0.91(avg). The hybrid RF-SVM classifier produces kappa measures of 93.1%, 91.87%, 89.92 and 91.04% against experimental, FSDD, ELSDSR and TIMIT datasets. The observed results show that proposed RF-SVM classifier method produces a higher kappametric and f1_score (avgof92%) value and thus offers improved speaker identification. It is not worthy that the f1_score of most systems drops sharply when using various datasets, which is likely due to speaker data imbalance. The hybrid RF-SVM classifier yields higher performance measures (accuracy, kappa and f1_score) than the other state-of-the-art systems using spectral features and improves the performance of the systems overall.

V CONCLUSION

The goal of this research is to provide a system for categorizing the severity of vocal pathology problems in order to facilitate more effective treatment. Silence removal and noise reduction are possible steps after the initial processing of various speech signals, such as the input of a vocal pathology condition. The Electro Grotto Graph (EGG) was used in conjunction with the Wiener and DWT filters to eliminate them. Voice designing and pathologic voice prediction are two applications of the Hybrid Wiener Filter Discrete Wavelet Transforms (HWFDWT) technique. It is suggested that the most important data be extracted from the exhibited pathologic speech sounds using a technique called Cat Swarm Optimization and Mel Frequency Cestrum Coefficients (CSOMFCC).

It has been shown that cautious feature selection and extraction may reduce both effort and data volume. The CSO algorithm improves the accuracy with which sex and voice abnormalities may be identified. After voice samples have been used for purposes like testing and teaching, their individual qualities are noted and stored in a database. The results of a CSOMFCC (Cat Swarm Optimization and Mel Frequency Cestrum Coefficients) study shed more light than those of MFCC and LPC studies. Using the proposed approach CSOMFCC, the features were recovered in reduced dimensionality and in a shorter amount of time.

To correctly label audio signals, a Back Propagation Neural Network (BPNN) is employed. To determine if a given speech sample is normal or abnormal, a computerized Automatic Speech Aberrant Identification System is used during the categorization phase. In order to make this phone call, you'll need to use the MOBPNDVC (Modified Optimized BPNC Disturbance Voice Classification). Both healthy and sick sounds can be identified by MOBPNDVC, and the system can randomly generate male and female sick voices.

As a result, ROC curves are constructed for a wide variety of conditions and tools, such as laryngitis, diplophonia, dysphonic, laryngoscopes, and chordates. An adaptation of the Support Vector Machine (SVM) for use in voice classification was inspired by the Modified Optimized Back Propagation Network Disorder Voice Classification (MOBPNDVC). By utilizing the widely-known and modified Back propagation neural network models, the proposed MOBPNDVC was able to achieve an accuracy of 97.50% on the Real-time Dataset of the Department of Pathology at the Karpagam Faculty of Medical Sciences and Research in Coimbatore, and 977.90% on the Saarbrucken data set test model. This paper presents and analyses the results of our investigation into the potential of an AVC system for identifying abnormal speech patterns.

REFERENCES:

1. *Fant, G 1960, Acoustic Theory of Speech Production (Mouton, the Hague, the Netherlands), Google Scholar, pp. 169-185.*
2. *Fant, G, Liljencrants, J & Lin, Q 1985, 'A four-parameter model of glottal flow', STL-QPSR, vol. 26, no. 4, pp. 1-13.*
3. *Faycal, Y & Messaoud, B 2014, 'Comparative performance study of several features for voiced/unvoiced classification', The International Arab Journal of Information and Technology, vol. 11, no. 3, pp. 293-299.*

4. *Feng, L 2004, 'Speaker recognition, informatics and mathematical modeling, technical university of denmark, DTU', English Language Speech Database for Speaker Recognition (ELSDSR) [Dataset], Available from: <http://www2.imm.dtu.dk/~lfen/elsdsr/>.*

5. *Ferrer, L, Lei, Y, McLaren, M & Scheffer, N 2016, 'Study of senone-based deep neural network approaches for spoken language recognition',IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 105-116, Available from:<https://doi.org/10.1109/ TASLP.2015.2496226>.*

6. *Fine, S, Navratil, J & Gopinath, R 2001, 'A hybrid GMM/SVM approach to speaker identification', Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001), Salt Lake City,Utah, vol. 1, no. 1, pp. 417-420, DOI: 10.1109/ICASSP.2001.940856.*

7. *Flanagan, JL 1972, 'The mechanism of speech production', in Speech Analysis Synthesis and Perception, Kommunikation and Kybernetik in Einzeldarstellugen, Springer, Berlin, Heidelberg, vol. 3, Available from: <https://doi.org/10.1007/978-3-662-01562-9>.*

8. *Freund, Y 1995, 'Boosting a weak learning algorithm by majority', Information and Computation, vol. 121, no. 2, pp. 256-285, Available from: <https://doi.org/10.1006/inco.1995.1136>.*

9. *Freund, Y & Schapire, RE 1997, 'A decision-theoretic generalization of on-line learning and an application to boosting', Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, Available from: <https://doi. org/10.1006/jcss.1997.1504>.*

10. *Garcia-Romero, D, Zhang, X, McCree, A & Povey, D 2014, 'Improving speaker recognition performance in the domain adaptation challenge using deep neural networks', IEEE Spoken Language Technology Workshop (SLT), pp. 378-383, Available from: <https://doi.org/10.1109/SLT.2014. 7078604>.*

11. *Ge, Z, Iyer, AN, Cheluvaraja, S, Sundaram, R & Ganapathiraju, A 2017, 'Neural network based speaker classification and verification systems with enhanced features', in Intelligent Systems Conference Intel. Sys, IEEE, pp. 1089-1094, DOI:10.1109/INTELLISYS.2017.8324265.*