A Comparison of Machine Learning Methods using Correlated Speech Features in the Presence of Varied Noise

D. U. R Khan[1†], Syed Abbas Ali[2††], and Hina D. Khan[3†††],

NED University of Engineering and Technology, Karachi, Pakistan

Abstract
Speech signal analysis processing helps extract information from both clean and noisy speech signals, and machine learning algorithms provide robust analytical tools for signal exploration. In this research (14) speech signal features were analyzed using machine learning tools with the following corpuses of speech commands: clean speech, with average noise, and with high noise. The analysis is based on the selection of the most correlated feature of distant and noisy speech along with the implementation of three (03) conventional learning (random forest nearest neighbor, voting model, and support vector machine (SVM)) and deep learning (Long short-term memory (LSTM)) models. This study presents a comprehensive result of selected features with clean, average noise, and very high-noise speech corpuses. The respective signal features performed well with a support vector machine (SVM) with no noise and average noise corpuses. However, LSTM shows significant results with high-noise corpus inters with macro-and average-weighted accuracy.

Keywords:
learning algorithms, LSTM, robust speech, speech.

1. Introduction

In this era of technology, speech has been transformed through the interaction and interface of machines. Many speech recognition systems have obtained satisfactory results, although many features of speech recognition exhibit different results in clean and noisy environments. There are many methods to achieve human-machine interfaces, and distant speech recognition (DSR) is considered the most genuine among them. Examples of DSR are commonly seen in Google Home, smart TVs, and Amazon Echo, which clearly show that it uses distant microphones for speech detection. There are many obstacles that make it difficult to build a solid distant speech recognition system, including overlapping speakers and background noise. The purpose of this study is to examine all aspects and features of the speech-by-speech processing and analysis process, and then to apply machine learning (random forest nearest neighbor, support vector machine, and voting model) and deep learning (convolutional neural network and LSTM), and to compare the features and results of all models of speech with and without noise and distance [8,13,20]. In addition, speech features were analyzed to enhance the performance of distant robust speech recognition and compare the extracted and selected feature performances with average and high noise.

The preceding feature extraction investigations are discussed in Section 2. The steps for speech signal processing and remote speech feature extraction and selection are outlined in section 3. The results are described in Section 4. On the basis of the data obtained, Section 5 concludes the performance of various learning algorithms in the presence of varied noise.

2. Literature Review

There are many different techniques used in distant speech recognition systems that are dependent on the division or type of DSR. The two distinct divisions of DSR are (i) a front-end speech augmentation system and (ii) a back-end automated speech recognition (ASR) system. Both studies used single or several distant microphones for voice recording. Advanced front-end microphone array techniques were applied to a DSR system with multiple distant microphones. The results imply that a lower word error rate (WER) is obtained compared to that condition when a single microphone is used. A large rift exists between automatic speech recognition and acoustic array processing, although significant progress has been made in both dimensions. In many cases, groundbreaking progress is to be made in the emerging field of DSR, and this abysmal state of affairs must change. This study outlines five pressing

Manuscript received April 5, 2022
Manuscript revised April 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.4.63

problems in DSR, a research field that is essential for constructing truly effective DSR systems [12,14]. The time-domain waveform contains all the auditory information of the speech signal. Various approaches for transforming data into information have been meaningfully interpreted in previous studies. To obtain statistically relevant data, audio signals must be converted into a small number of characteristics or features. Therefore, it is necessary to develop techniques for reducing information from incoming data. These features were categorized into segments and similar segments were grouped and compared. In terms of parameters, there are a variety of innovative and unique approaches for quantifying speech signals. Although they all have advantages and disadvantages, we have listed some of the more popular approaches along with their significance [1]. In most back-end state-of-the-art ASR systems used in distant speech recognition systems, the recognition problem is divided into three sub-processes: (i) extraction of features, (ii) acoustic modelling, and (iii) language modelling. To obtain the best performance, each was refined independently. Discriminative characteristics are obtained from speech signals by feature extraction to classify linguistic content, which leads to obtaining perceptual linear prediction coefficients (PLPs) and Mel-filter bank cepstral coefficients (MFCCs). Based on these results, many speech-related systems gain optimal efficiency [2]. Speech is a profound and natural human skill. The coordination of approximately 100 muscles and 14 different sounds/s was mainly characterized in grownups. Speaker identification from speech signals is mainly dependent on the hardware or software capacity to detect speech signals and then identify the speaker in it. [3] Noise (natural or distant multiple speakers) must be removed by treating speech before feature extraction and speaker identification [4,9]. A preset amount of the signal component is used to visualize a voice signal, which is counted as the goal of feature extraction. It may be time-consuming if we treat all the data in the acoustic signal, as it is irrelevant in the identification task. [5,6].

3. Methodology

The distant speech recognition approach is a technology that relies on traditional speech signal analysis techniques to analyze the features that are most effective in speech recognition, as seen in the light of previous datasets and research. Distant speech recognition and analysis include preliminary processing for speech signal processing, feature extraction, and classification by machine learning and deep learning user-defined ensemble modelling [21]. This study selects a speech signal analysis approach with a new user-defined algorithm that contains the speech processing, feature extraction, and selection process with user-defined ensemble learning of machine learning and deep learning models.

A model was built for each algorithm and tested using a data sample from the TensorFlow speech recognition dataset. Fourteen speech features are included in the process of feature extraction, root mean square (RMS), Chroma variant "Chroma Energy Normalized" (CENS), roll-off frequency, poly features, Mel-scaled spectrogram, Zero passage crossing (ZPR), Spectral contrast, Mel-frequency cepstral coefficients (MFCCs), Spectral centroid, Relative spectral perceptual linear prediction (Rasta-PLP), Short-Time Fourier Transform, Chroma Features (Chroma STFT), Linear predictive cepstral coefficients (LPCC), Tonal centroid features (tonnetz) and Pitch [19]. For feature selection, we applied the correlation method to analyze the most correlated and influential features in distant speech recognition. Feature data transformation and preparation includes reshaping feature vectors and merging them into single sample vector for each sample in training and testing and then we saved each sample into the data frame and them we split that data frame into training and testing data then encoded word labels of each sample, after all above steps we configured and trained all models which are used in our algorithm and then ensemble the testing step of each model to create ensemble function for final results this processes includes Random forest, K nearest neighbors, Convolutional neural network, Support vector model, from machine learning 'Voting classifier' and Long short term memory (LSTM) model for classification of features and signal . The steps of the speech signal analysis process implemented for distant speech feature extraction and selection are shown in Figure.1.

A. Distant Robust Speech Dataset collection (Clean speech, With average noise, with very high noise)
B. Speech Signal Acquisition
C. Speech Signal pre-processing
D. Feature Extraction
E. Feature Selection and Analysis
F. Feature Data transformation and Preparation

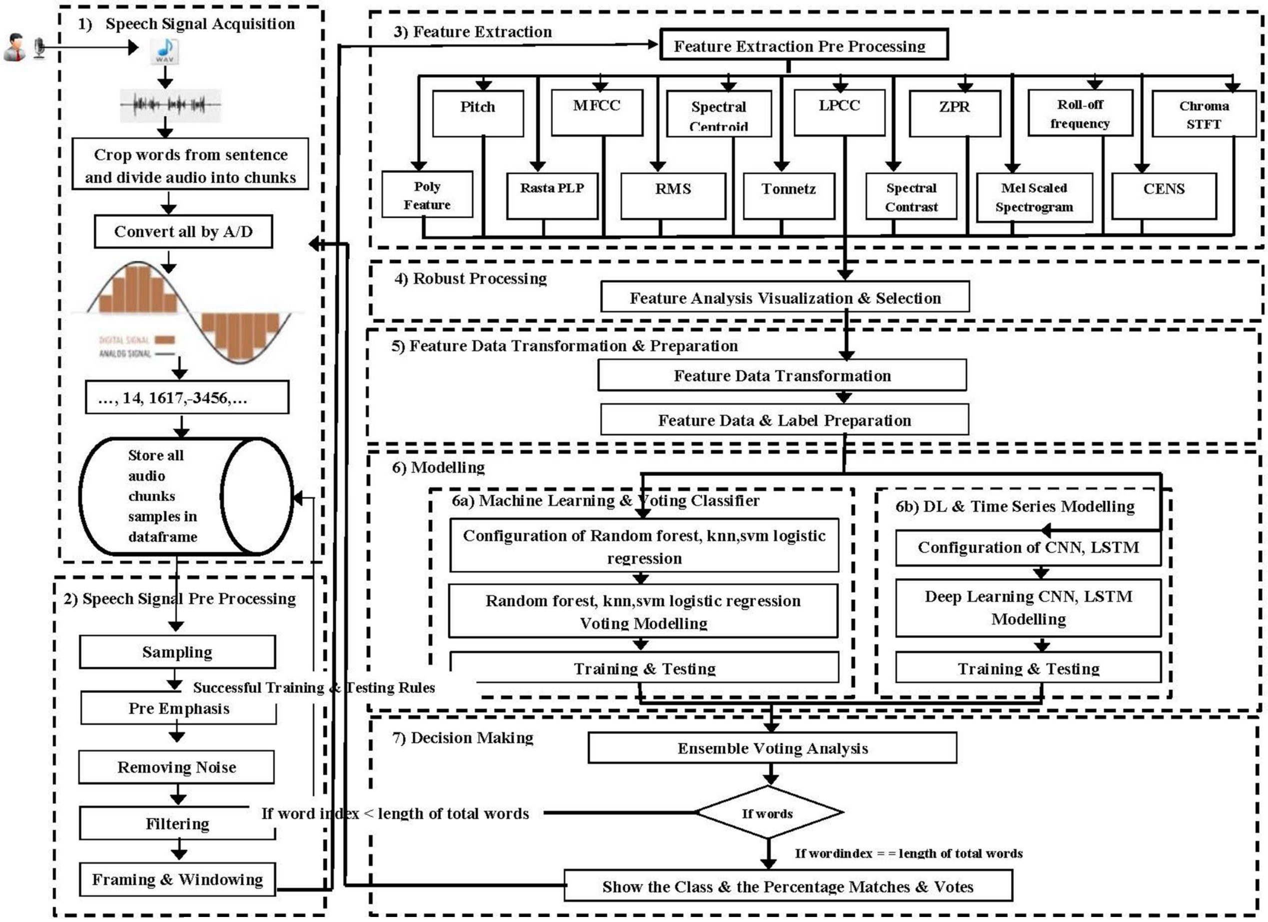

Figure 1: Distant Robust Speech Feature Extraction and Analysis Flow

A. Distant Robust Speech Dataset Collection (Without Noise, With Average Noise, With Very High Noise)

The primary source of the dataset was the TensorFlow speech-recognition dataset. See Table 1. Data were collected in the form of audio samples. The most recent speech command dataset is released by TensorFlow, which includes 30 short words with 65,000 one-second-long utterances by thousands of different people. The sample dataset is split into two portions: test and training datasets (the training dataset is 70% and the test dataset is 30%). The dataset contained 1000+ wav format audio recording files for each word.

B. Speech Acquisition

Inaccurate or unreliable records from the database or dataset are detected and removed/corrected during the process of speech data cleaning. After the detection of the non-relevant part of the dataset, it is remodeled or the unwanted dirty data are removed. This process is applied to batch processing through interactive scripting using data wrangling tools.

C. Speech Pre-Processing

To extend the potency of the desired feature, the sound preprocessing stage is applied in the sound recognition system, and recognition performance is boosted in the classification stage. Speech preprocessing includes three major steps: dataset sampling, windowing, and noise removal. Sampling was performed to obtain a discrete signal from a time continuous sound signal, and the result was a time- and value discrete signal *x(t)*. To obtain a finite frequency f_{max}, it is band-limited, and a sampling frequency is used to sample at least $2f_{max}$. In this manner, it can be reconstructed using its time-discrete signal *x[n]*. A direct impact on recognition accuracy was demonstrated by Sanderson et al. [7]

Words	Without Noise	With Average Noise	With High Noise
bed	1713	1356	1367
bird	1731	1357	1366
cat	1733	1424	1402
dog	1746	1498	1489
down	2359	1198	1223
eight	2352	1133	1113
five	2357	1092	1112
go	2372	960	960
four	2372	2400	2400
happy	1742	1481	1481
left	2353	1505	1485
house	1750	2392	2392
marvel	1746	1253	1253
nine	2365	1144	1145
no	1881	1002	963
off	2357	2244	2252
on	2392	2228	2228
one	2370	1276	1276
right	2367	1296	1276
seven	2392	1411	1411
six	2387	1485	1500
sheila	1734	1463	1463
three	2346	1188	2028
stop	2390	1485	1485
tree	1733	1188	2072
two	2373	902	902
up	2376	1187	1187
wow	1745	957	957
yes	2387	1244	1547
zero	2376	1306	1602

Table 1: Dataset Information

D. Speech Features Extraction

This study focuses upon the following selected fourteen features of speech recognition.

1) Mel-frequency cepstral coefficients (MFCCs)

Mel frequency cepstral (MFC) can be considered as a collection of Mel-frequency cepstral coefficients (MFCCs), which are derived from the audio clips' cepstral representation. It can simply be defined as a 'non-linear spectrum-of-a-spectrum' i.e. short phase of power spectrum of any audio or sound signal. A type of inverse Fourier transform (cepstral) representation can be used to derive this equation. The MFC allows for a more accurate depiction of sound because the frequency bands are evenly dispersed on the Mel scale, closely resembling the response of the human auditory system.

2) RMS value of every frame

RMS is abbreviated as 'root-mean-square', and is defined as the square of the amplitude of each wave form over one complete cycle; the average is taken as the sum of the values of the overall signal. At each spectral line, the RMS amplitude format was used to describe the comparable steady-state value of the sine wave. RMS of a spectrum: It is desirable to determine the RMS value of a spectrum. The RMS of a spectrum is a single value that represents the overall amount of energy present across the frequency range.

3) Chroma variant "Chroma Energy Normalized" (CENS)

The term chromagram refers to putting all pitches in an audio recording in one location so that we can understand how to classify them. It is a metric for measuring sound quality that allows one to categorize sounds as higher, lower, or medium. CENS features work by smoothing local irregularities in speed, articulation, and melodic ornaments such as trills and arpeggiated chords using statistics over vast windows. CENS is best used for speech audio matching and similarity tasks [10].

4) Mel-scaled spectrogram

The output of a nonlinear frequency scale translation is the Mel scale. The sounds are equal in distance from one another in Mel scale measures, which contain a set of pitches that the listener perceives as equal in distance. The gap between 7500 and 8000 Hz is barely discernible on the Mel-scale compared to the Hz scale, where a clear difference is found between 500 and 1000 Hz [11].

5) Spectral Centroid

The spectral centroid of a signal is the curve whose value at any given time corresponds to the centroid of the spectrogram's associated constant-time cross-section. A noise resistant assessment of how a signal's main frequency varies over time is provided by the spectral centroid. The location of the center of mass of the spectrum is known as the spectral centroid. The spectral centroid is a measure that can be useful in characterizing the spectrum of an audio file signal because audio files are digital signals. This is sometimes referred to as the spectrum's median; however, there is a distinction between the spectral centroid and the spectrum's median measurement.

6) Tonal Centroid features (tonnetz)

Tonnetz, which illustrates single-step voice-leading relationships among major and minor triads and was essential in Richard Cohn's early papers, is our fundamental example of a note-based graph (see esp. Cohn 1996, 1997) [16].

7) Spectral Contrast

The change in the sound energy distribution over frequency is represented by spectral contrast. These individuals are likely to have problems with neuronal processing that uses spectral contrast to minimize noise. A class of methods that achieve global spectral contrast enhancement has resulted from research on artificial compensation for spectral contrast deficits. The spectral contrast is a measurement of the frequency energy at each timestamp.

8) Nth order Polynomial (Poly) Feature

Approximations of local polynomials have a versatile feature space, particularly in time-domain signal analysis. In addition, 'efficient recursions' and 'autonomous linear statespace models are utilized to calculate the parameters of such polynomials, which leads to effective analytical solutions.

9) Frequency Roll-off

In many networks, roll-off builds a continuous gradient at frequencies considerably beyond the cutoff point of the frequency. Frequency roll-off is implied in the respective study to lower the cutoff performance to a single number of such a filter network.

10) Short-Time Fourier Transform and Chroma Features (Chroma STFT)

STFT Chroma, the strength of the 12 separate pitch classes used to study music, is represented by the chroma value of an audio. These can be used to distinguish between different pitch-class patterns in audio sources.

11) Zero Crossing Rate (ZCR)

The rate at which the sign of the signal changes within the frame of an audio is known as the zero-crossing rate (ZCR). In simpler words, the number of times the signal is switched from +ve to –ve and back divided by the length of the frame is ZCR. Technically, this means that it is the rate at which the signal changes from positive to negative and negative to positive, or the rate at which the signal crosses the zeroth line [15].

12) Linear predictive cepstral coefficients (LPCC)

Linear prediction cepstral coefficients (LPCC) are the coefficients that are produced from the LPC-computed spectral envelope (see esp. Alim, Rashid 2018) [21]. In technical terms, this is the coefficient of the logarithmic magnitude of the LPC spectrum, as visualized from the Fourier transform.

13) Relative spectral-perceptual linear prediction (Rasta-PLP)

The spectral transform was used in the Rasta PLP perceptual linear prediction. The method of wrapping spectra to reduce variations between speakers was introduced by Hermansky. In this method, vital speech information is retained, which aids in linear predictions [17]. In this technique, short-term noise is smoothed by passing it through a band-pass filter, and static spectral coloring in the audio channel removes continuous offset [18].

14) Pitch

The fundamental period of a spoken signal is called the pitch. The fundamental frequency has a perceptual relationship. It depicts the vibration frequency of vocal cords during sound generation (e.g., vowels). In speech, pitch refers to the perceived highness or lowness of a tone, which is determined by the number of vibrations per second generated by vocal cords.

E. Feature Selection

To decrease the number of features, a mixed manual code selection technique was employed to select the most important features. To make the analytical procedure easier, quantitative or continuous features were chosen from among the 14 explanatory features. These features were chosen with great care. The desired columns are kept, while the unwanted columns are removed.

The correlation matrix of feature features is shown in Figure 2, with highly correlated features shown in white.

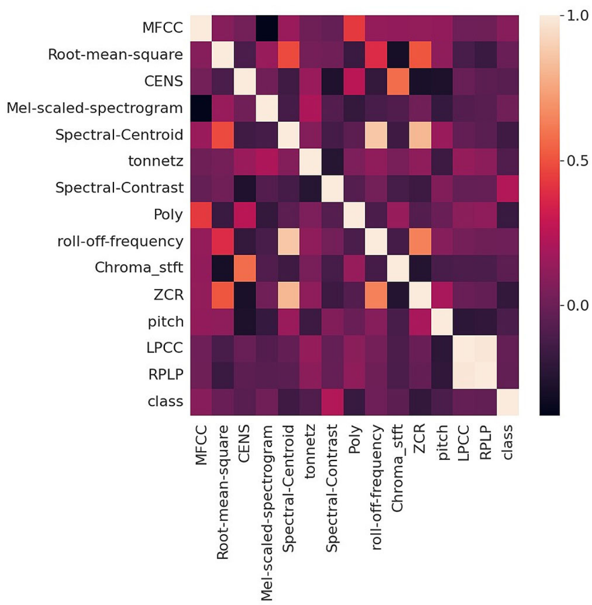

Figure 2: Correlation matrix heat map for feature selection

As a result, the four features chosen to be the explanatory or independent features that will operate as predictors of the response or dependent feature of the class of words picked by a correlation matrix are MFCC, Mel-scaled spectrogram, Poly feature, and Zero crossing rate. To satisfy the assumption of no collinearity between the independent features, a correlation matrix analysis was conducted, and the results are shown in Table. 2.

S.No	Features	Information	Type
1.	MFCC	0.139776	Numeric
2.	Mel-scaled-spectrogram	0.219794	Numeric
3.	Poly Feature	0.146756	Numeric
4.	ZCR	0.120788	Numeric
5.	class	1.000000	Numeric

Table 2: Most Correlated features by correlation

In Table. 2, the selected features that have a high correlation rate are selected as feature sets for random forest, SVM, k-nearest neighbors, voting classifier, and LSTM modelling.

	MFCC	RMS
0	[-507.53931378 -554.92349096 -737.54283489	[0.00041371 0.00041302 0.00032775 ... 0. ...
1	[-661.36467817 -594.74552722 -516.87036514	[0.00071286 0.00392151 0.00485728 ... 0. ...
2	[-651.79179295 -574.46392933 -551.33330276	[0.0039245 0.00343906 0.00348912 ... 0. ...
3	[-502.47100708 -527.32817489 -602.64260131	[0.00328921 0.00322399 0.00332264 ... 0. ...
4	[-566.95015514 -597.92443726 -690.05951258	[0.00050448 0.00051213 0.00050514 ... 0. ...

	s_centroid	tonnetz
0	[2698.28283671 2677.42594362 2869.13835581	[-0.04073299 -0.05662396 -0.00869005 ... 0. ...
1	[2592.35034917 3967.06907565 4919.71108822	[-0.03114112 0.01321885 -0.0485156 ... 0. ...
2	[4519.40944803 4994.59665081 5158.26461524	[-0.04333293 0.02727674 -0.03083694 ... 0. ...
3	[2818.28315562 2745.85049777 2717.020046	[0.04000741 -0.03768926 -0.07043758 ... 0. ...
4	[2721.27361512 2686.03854976 2733.62417348	[0.13270239 -0.02998883 -0.09436941 ... 0. ...

	s_contrast	poly
0	[0.08928246 0.08927364 1.12494679 ... 0. ...	[-3.11258490e-06 -1.92588490e-06 -7.66799186e-...
1	[0.12676975 0.12689537 1.28289219 ... 0. ...	[-2.41970194e-06 -1.31485757e-07 1.71519186e-...
2	[0.05633728 0.05560613 6.73914048 ... 0. ...	[1.41784535e-06 1.29625451e-05 2.22098452e-05 ...
3	[0.05221869 0.052181 2.79744679 ... 0. ...	[-9.65395335e-06 -1.14392351e-05 -1.15659411e-...
4	[0.09467778 0.09466945 1.33508297 ... 0. ...	[-2.08513852e-06 -2.04941875e-06 -1.79967690e-...

	ROF	CS
0	[7125. 6976.5625 6437.5 ... 0. ...	[0.79139468 0.69036111 0.55793248 ... 0. ...
1	[3750. 7250. 7437.5 ... 0. 0. 0.]	[0.10586217 0.53820438 0.33082177 ... 0. ...
2	[7390.625 7445.3125 7445.3125 ... 0. ...	[0.22640194 0.27796745 0.25796955 ... 0. ...
3	[4750. 3820.3125 3765.625 ... 0. ...	[0.67808504 0.37168226 0.24937521 ... 0. ...
4	[6023.4375 5062.5 4382.8125 ... 0. ...	[0.22044174 0.28858153 0.37211326 ... 0. ...

	ZCR	pitch
0	[0.16259766 0.24462891 0.32861328 ... 0. ...	0 0.0\n1 0.0\n2 0.0\n3 ...
1	[0.15771484 0.29882812 0.46484375 ... 0. ...	0 0.0\n1 0.0\n2 0.0\n3 ...
2	[0.24755859 0.42236328 0.54833984 ... 0. ...	0 0.0\n1 0.0\n2 0.0\n3 ...
3	[0.18115234 0.27001953 0.36181641 ... 0. ...	0 0.0\n1 0.0\n2 0.0\n3 ...
4	[0.17578125 0.26074219 0.34033203 ... 0. ...	0 0.0\n1 0.0\n2 0.0\n3 ...

	LPCC	RPLP	Class
0	[-2.86036292 -1.47283167 -0.25077087 ... 0. ...	[-0.65587032 -0.39933653 -0.26280394 ... 0. ...	tree
1	[-2.86036292 -1.47283167 -0.25077087 ... 0. ...	[-0.65587032 -0.39933653 -0.26280394 ... 0. ...	tree
2	[-2.86036292 -1.47283167 -0.25077087 ... 0. ...	[-0.65587032 -0.39933653 -0.26280394 ... 0. ...	tree
3	[-2.86036292 -1.47283167 -0.25077087 ... 0. ...	[-0.65587032 -0.39933653 -0.26280394 ... 0. ...	tree
4	[-2.86036292 -1.47283167 -0.25077087 ... 0. ...	[-0.65587032 -0.39933653 -0.26280394 ... 0. ...	tree

Figure 3: Feature data before transformation

	MFCC	MSS
0	[-526.11171609 -546.97386513 -642.36240536	0 6.726135e-06\n1 1.681460e-06\n2 ...
1	[-546.71911121 -520.31038094 -530.84725632	0 1.052596e-07\n1 2.631331e-08\n2 ...
2	[-466.69731192 -502.1146657 -648.64455083	0 5.716309e-05\n1 1.429023e-05\n2 ...
3	[-335.0624734 -354.33016538 -462.13864609	0 1.072225e-04\n1 2.680514e-05\n2 ...
4	[-362.21780626 -397.67210147 -516.77317309	0 5.230556e-04\n1 1.307584e-04\n2 ...

poly	ZCR	Class
[-6.19621342e-06 -7.28602576e-06 -6.83863206e-...	[0.15820312 0.2421875 0.32226562 ... 0. ...	bed
[-1.36422484e-06 -6.64285797e-06 -1.25245452e-...	[0.15185547 0.23144531 0.31347656 ... 0. ...	bed
[-5.35001109e-06 -5.09831532e-06 -4.45097004e-...	[0.15136719 0.21826172 0.29833984 ... 0. ...	bed
[-3.55696557e-05 -3.97034256e-05 -3.86694552e-...	[0.18212891 0.27099609 0.35888672 ... 0. ...	bed
[-2.04740496e-05 -1.90383202e-05 -1.38347846e-...	[0.13671875 0.20263672 0.27197266 ... 0. ...	bed

Figure 4: Feature data after transformation

F. Feature Data Transformation and Preparation

Feature transformation is a group of techniques that creates new features (predictor features). Feature selection is a subset of feature transformation it is done for Knowledge discovery, Interpretability, to gain some insights and Curse of dimensionality, there are two ways of feature selection. Filter type techniques select features regardless of the model one of them correlation method was applied and explained in above heading. It mainly depends on the general features that help in prediction. Filter techniques suppress the least interesting features. They are mainly applied as preprocessing methods. A subset of features can be evaluated by the application of Wrapper techniques, which allows, unlike filter approaches, the detection of possible interactions between features. You can see the data frame in Figure. 3, in which data was stored and only the above four selected feature columns were selected for the final transformation, as shown in Figure. 4.

4. Discussion and Obtained Results

The Table. 3 shows a comparison of the features of clean speech, with average noise, and with high noise speech. MFCC, Mel-scaled spectrogram, Poly feature, and zero crossing rate are more correlated and effective as features of noisy speech recognition. Three machine learning predictive models were studied in their respective papers: random forest, K-nearest neighbors, support vector machine, ensemble learning model voting classification, and deep learning model LSTM. Predictive approaches are likely appropriate for situations involving complexity and uncertainty. They will be very useful, although it is sometimes very difficult to model problems in such a way that they could be more convincing than practically useful. The deep learning model is simpler than the machine learning model in terms of the mathematical equations that constitute the computational nature the models are trying to solve.

Features per interval	Without Noise	With Average Noise	With High Noise
MFCC	Low intense energies	Medium intense energies	High intense energies
RMS value for each frame	High in range of 10^{-1} to 10^{-2}	Low in range of 10^{-2} to 10^{-3}	Low in range of 10^{-2} to 10^{-3}
Chroma variant CENS	Low intense energies	Medium intense energies	High intense energies
Mel-scaled spectro-gram	Low additive noise	Medium additive noise	High additive noise
Spectral Centroid	Low 2400 to 3400db	High 2500 to 5000db	High 2500 to 5000db
Tonal Centroid features (tonnetz)	6 pitch classes high pitch more between 0 to +1 and less between 0 to -1	6 pitch classes medium pitch between 0 to -1 and high between 0 to +1	6 pitch classes high pitch low between 0 to -1 and high between 0 to +1
Spectral Contrast	Low contrast	Medium contrast	High contrast
Poly Feature	High between 0 to 1.4Hz	Medium between 0 to 0.5	Low between 0 to 0.3
Roll-off frequency	Low 4200 to 5900 Hz	Medium 4000 to 7500Hz	High 4000 to 7500 Hz
Chroma stft	Low intense energies	Medium intense energies	High intense energies
Zero Crossing Rate	Low Zero crossing	Medium Zero crossing	High Zero crossing
LPCC	Low intense energies	Low intense energies	Low intense energies
Rasta PLP	Low intense energies	Low intense energies	Low intense energies
Pitch	Low intense pitch	Medium intense pitch	High intense pitch

Table 3: Comparison of features according to without n

When it is implemented in statistical software, which is in this case Python, the deep learning results are much easier to interpret and more accurate than machine learning. For instance, the comparison may be related to the predicted values and actual values of the test data. The predicted values generated by the model based on the same test data were provided by both deep learning and machine learning. Based on these two array values, actual and predicted, we can assess the evaluation criterion of the root mean square error (RMSE), Pearson correlation coefficient R, and mean absolute error (MAE). These comparison metrics or criteria were defined using the following mathematical formulae(1),(2), and (3):

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n} |(y_i - \hat{y}_i)| \quad (1)$$

where yi is the~~,~~ actual value, yi is represented as $\boldsymbol{y_i}$ and predicted value is presented as $\boldsymbol{\hat{y}_i}$ and $\boldsymbol{n}$ is the number of observations.

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |(y_i - \hat{y}_i)|^2} \quad (2)$$

$$\boldsymbol{R} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y} - \bar{\hat{y}})^2}} \quad (3)$$

Where $\boldsymbol{y_i}$ is the actual value and $\boldsymbol{\hat{y}_i}$ is the predicted value, $\boldsymbol{\bar{y}}$ is the mean of actual value, $\boldsymbol{\bar{\hat{y}}_i}$ is the mean of predicted value and $\boldsymbol{n}$ is the number of observations. Figures 5, 6, and 7 depict the prediction performance of both models, while Tables 4, 5, and 6 provide information for the three-comparison criterion described above. Compares the performance of all algorithms with the chosen features.

Criterion	RF	KNN	SVM	Voting	LSTM
MAA	0.95	0.80	0.97	0.91	0.94
WAA	0.95	0.80	0.97	0.91	0.93
Score	0.85	0.62	0.93	0.80	0.93
MAE	0.76	1.99	0.327	0.99	0.79
RMSE	3.0	4.41	2.0	3.45	3.5
MSE	12.0	28.0	5.01	13.4	12.3
PCC	0.95	0.82	0.97	0.91	0.94
P value	0.0	0.0	0.0	0.0	0.0

Table 4: Comparison Indicators Without Noise

Criterion	RF	KNN	SVM	Voting	LSTM
MAA	0.93	0.79	0.96	0.90	0.93
WAA	0.93	0.79	0.96	0.90	0.93
Score	0.83	0.60	0.92	0.78	0.92
MAE	0.78	2.08	0.427	1.06	0.76
RMSE	3.52	5.41	2.43	4.00	3.3
MSE	12.44	29.28	5.914	16.07	12.4
PCC	0.91	0.81	0.96	0.89	0.93
P value	0.0	0.0	0.0	0.0	0.0

Table 5: Comparison Indicators Average Noise

Criterion	RF	KNN	SVM	Voting	LSTM
MAA	0.87	0.66	0.91	0.83	0.93
WAA	0.87	0.67	0.91	0.83	0.93
Score	0.76	0.34	0.87	0.628	0.92
MAE	1.28	3.42	0.69	1.868	0.76
RMSE	4.21	7.10	3.053	5.329	3.5
MSE	17.7	50.4	9.32	28.40	12.3
PCC	0.88	0.68	0.939	0.829	0.93
P value	0.0	1.91	0.0	0.0	0.0

Table 6: Comparison Indicators With High Noise

Where, PCC stands for Pearson correlation coefficient, MAA stands for Macro Average Accuracy, and WAA stands for Weighted Average Accuracy.

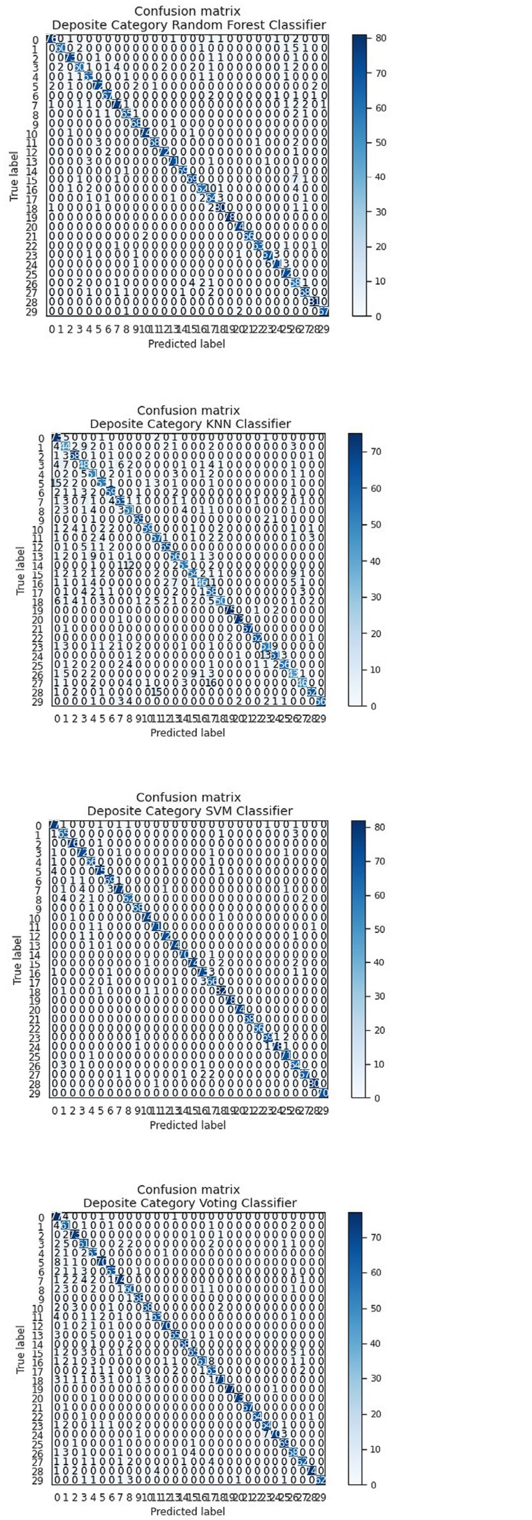

Figure 5: Actual versus predicted of test data Confusion Matrices of No Noise by Machine Learning and Deep learning with selected Features MFCC, Mel scaled spectrogram, Poly feature and Zero Crossing rate

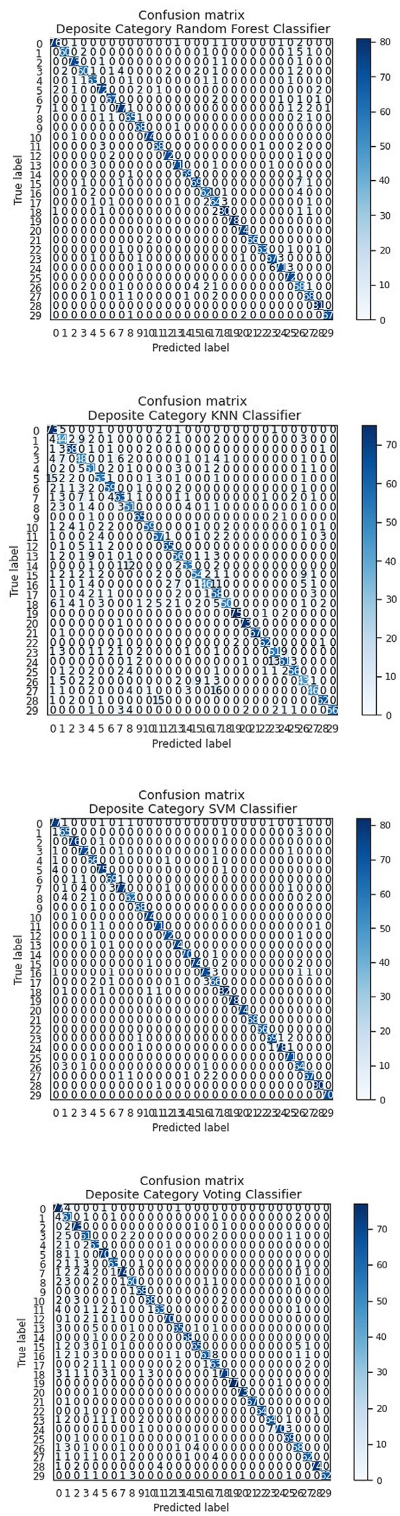

Figure 6: Actual versus predicted of test data Confusion Matrices of Average Noise by Machine Learning and Deep learning with selected Features MFCC, Mel scaled spectrogram, Poly feature and Zero Crossing rate

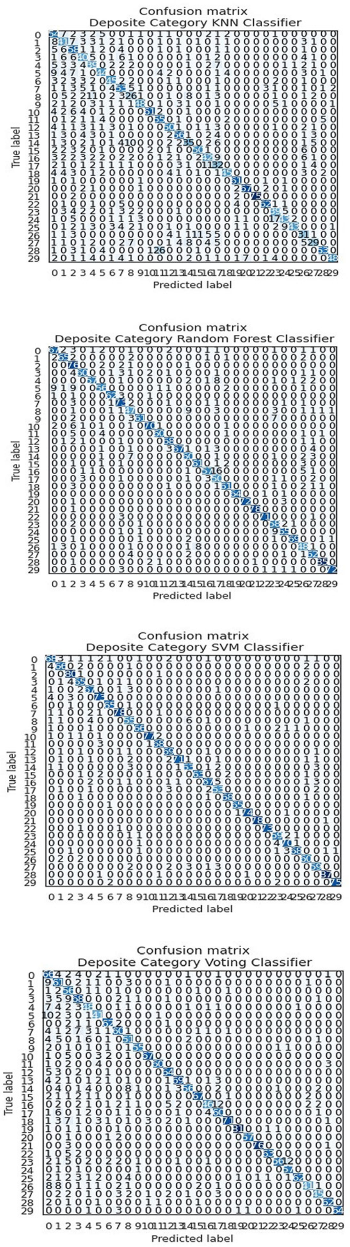

Figure 7: Actual versus predicted of test data Confusion Matrices of High Noise by Machine Learning and Deep learning with selected Features MFCC, Mel scaled spectrogram, Poly feature and Zero Crossing rate

Table. 4, 5, and 6 shows that different selected features like MFCC, Poly, Mel scaled Spectrogram and Zero crossing rate outperform with Support vector machine in comparison with other algorithms. Whereas, SVM shows least values of RMSE and MAE and MSE in the absence of noise. SVM also illustrate the significant values of SVM with RMSE and MAE and MSE in comparison with other algorithms with average noise. On the other hand, LSTM shows considerable results in comparison with SVM and other learning algorithms. can be seen with SVM algorithm It is also higher in Score, SVM comprises on higher values of Pearson correlation coefficient with no noise, average noise and high noise show strong correlation.

5. Conclusion

This study evaluated the performance of different learning algorithms in the presence of three different categories of noise (no, average, and high) with highly correlated distant speech features. In the experimental framework, data were collected from the tensor flow speech recognition corpus in the form of audio samples with different levels of noise. Experimental results show that the Support vector machine (SVM) outperform with highly correlated features in comparison with other learning algorithms with no noise and average noise speech sample in term of macro and average weighted accuracy. In contrast, LSTM performs well in comparison with other algorithms in the presence of a high-noise speech sample, with the lowest values of RMSE, MAE, and MSE.

References

[1] Lu, S., Li, Z., Qin, Z., Yang, X. and Goh, R.S.M., 2017, December. A hybrid regression technique for house prices prediction. In 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 319-323). IEEE.

[2] Brownlee, J., 2018. Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.

[3] De Cock, D., 2011. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3).

[4] Kant, S., 2010. Machine Learning and Pattern Recognition. Defence Science Journal, 60(4), p.345.

[5] Denil, M., Matheson, D. and De Freitas, N., 2014, January. Narrowing the gap: Random forests in theory and in practice. In International conference on machine learning (pp. 665-673). PMLR.

[6] Hastie, T., Tibshirani, R. and Wainwright, M., 2019. Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC.

[7] Sanderson, C. and Paliwal, K.K., 2003. Noise compensation in a person verification system using face and multiple speech features. Pattern recognition, 36(2), pp.293-302.

[8] S. Debnath, P. Roy, V. Justin and S. Naik, "Study of different feature extraction method for visual speech recognition," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402357.

[9] F. Y. Chun and C. H. Juan, "Based on CFC And Multi-Feature Combination Optimization of Speech Recognition Research," 2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT), 2021, pp. 591-595, doi: 10.1109/ISCIPT53667.2021.00126.

[10] Müller, Meinard & Ewert, Sebastian. (2011). Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features.. 215-220.

[11] Umesh, S. & Cohen, Leon & Nelson, Douglas. (1999). Fitting the Mel scale. 1. 217 - 220 vol.1. 10.1109/ICASSP.1999.758101.

[12] Wölfel, M. and McDonough, J., 2009. Distant speech recognition. John Wiley & Sons.

[13] Ravanelli, M., 2017. Deep learning for distant speech recognition. arXiv preprint arXiv: 1712.06086.

[14] Wang, L., Kitaoka, N. and Nakagawa, S., 2006. Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN. EURASIP Journal on Advances in Signal Processing, 2006, pp.1-11.

[15] Gouyon, Fabien & Pachet, Francois & Delerue, Olivier. (2002). On the Use of Zero-Crossing Rate for an Application of Classification of Percussive Sounds.

[16] Cohn, R., 1997. Neo-riemannian operations, parsimonious trichords, and their" tonnetz" representations. Journal of Music Theory, 41(1), pp.1-66.

[17] Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. the Journal of the Acoustical Society of America, 87(4), pp.1738-1752.

[18] Hermansky, H. and Morgan, N., 1994. RASTA processing of speech. IEEE transactions on speech and audio processing, 2(4), pp.578-589.

[19] Alim, S.A. and Rashid, N.K.A., 2018. Some commonly used speech feature extraction algorithms (pp. 2-19). IntechOpen.

[20] Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A.E.D., Jin, W. and Schuller, B., 2018. Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Transactions on Intelligent Systems and Technology (TIST), 9(5), pp.1-28.

[21] Alhamada, A.I., Khalifa, O.O. and Abdalla, A.H., 2020, December. Deep learning for environmentally robust speech recognition. In AIP Conference Proceedings (Vol. 2306, No. 1, p. 020025). AIP Publishing LLC.

Danish Ur Rehman Khan received his B.E. degree in Computer Systems Engineering in 2000 from NED University of Engineering & Technology, Karachi, Pakistan. He received his first M.Engg. degree in Electrical Engineering with majors in Telecommunications in 2007 and second M.Engg. degree in Computer & Information Systems Engineering with majors in Computer Networks & Performance Evaluation (ISP) in 2012 from NED University. He is presently enrolled as Ph.D scholar in NED University. Current research interests includes Human Behavior and Speech Recognition. He is currently working as Technical Assistant To The Vice Chancellor at NED University. He served as Senior Manager (Networks & Hardware) at IT Department of NED University from 2007 to 2018. He is the senior member of IEP since 2017, and is playing very active role for the betterment of Engineers of Pakistan. He is also the member of PEC since 2000.

Syed Abbas Ali received his B.E. and M.Engg. degrees in Computer Systems and Electrical Engineering from the NED University of Engineering and Technology, Karachi, Pakistan in 1999 and 2002 respectively. Also, he received his PhD degree in Computer Science from NED University of Engineering & Technology, Karachi, Pakistan, in 2015. His research interests include Machine Learning and Speech Recognition, Speech Emotion Recognition and Computational Intelligence. Since 2020, he is working as a PROFESSOR in Department of Computer & Information Systems Engineering, NED University. He is serving the said department since 2000. He has more than 50 publications in the field of Machine learning and speech Recognition. He is the member of Pakistan Engineering Council since 1998 and member IEP since 2021.

Hina Danish Khan received her B.E degree in Computer & Information Systems Engineering in 2005 from NED University of Engineering & Technology, Karachi, Pakistan. She received her M.Engg degree in Computer architecture and performance evaluation in 2007 from the same department. Her research interests are Artificial Intelligence and Speech Recognition. She is currently serving as ASSISTANT PROFESSOR in the Department of Computer & Information Systems Engineering. She has taught the Subjects of Artificial Intelligence, Neural Networks, Digital Design, Fault Diagnosis, Real Time Systems and many more since she is working in the department. She is member of Pakistan Engineering Council since 2005. She is also the member of Institute of Engineers Pakistan since 2019.