Non-Intrusive Speech Quality Assessment

Swapnil.K.Gore
School of Electronics and Electrical Engineering
MIT Academy of Engineering
Pune, India
skgore@mitaoe.ac.in

Anagha.P.Haral
School of Electronics and Electrical Engineering
MIT Academy of Engineering
Pune, India
apharal@mitaoe.ac.in

Piyush.P.Walde
School of Electronics and Electrical Engineering
MIT Academy of Engineering
Pune, India
ppwalde@mitaoe.ac.in

Vanshika.R.Sonekar
School of Electronics and Electrical Engineering
MIT Academy of Engineering
Pune, India
sonekarvr@mitaoe.ac.in

***Abstract*−The speech signal is considered as a fastest method of interaction for human beings. This reality has motivated to think of speech as a fast and effective method of communication and improve the speech quality to make it more convenient. Intrusive method of speech quality requires clean reference signal which is not effective in many telecommunication applications Non-Intrusive method evaluates quality of signal without using clean reference signal. For this purpose, Quality of speech is assessed using Machine Learning model and Non-Intrusive speech quality assessment is proposed by extracting the MFCC (Mel Cepstral Frequency Coefficient) Features of the audio signal. The main objective is to detect the quality of speech enhancement and speech systems.**

Keywords−Speech quality assessment, Machine learning, MFCC.

I. Introduction

Communication evolves every time around us, whether it may be through speech, video, cell phones or any other means of communication. Evaluating the speech quality will help us to improve and build up our understanding of regular communication. Whenever you listen to someone who is speaking, you are habitually assessing them based on your morality. Hence there is need to assess the speech quality accurately and reliably.

The speedy growth in use of speech processing machine learning algorithms in applications like multi-media and Telecommunication has raised the demand for evaluating quality of speech. In non-intrusive speech quality assessment model input will be provided as audio signal. The features of audio of signal are then extracted using MFCC (Mel cesptral frequency coefficient) technique.

The extracted features along with the subjective mean score (MOS) are used for building the dataset. This dataset is passed to Support Vector Machine (SVM) Classifier for training the model. Then the score of speech is obtained as an estimated quality of speech signal.

II. Related Works

The study represented in [1] uses a Non-intrusive objective speech quality estimation using features at single scale and multiple scales. The feature computation techniques such as Lion' auditory model, MFCC, LSF have been compared and implemented. Several effective auditory features are combined together for training of model using Gaussian Mixture Model for speech quality estimation.

In this work, it was observed that the correlation of Mean opinion score (MOS) improves significantly, which shows the efficacy of the technique. One of the first non-intrusive signal-

based model was proposed by Liang and Kubichek, where perception-based speaker independent speech parameters such as perceptual linear prediction coefficients (PLP) and perceptually weighted Bark spectrum are utilized for nonintrusive speech quality evaluation. They compared the degraded speech parameters to an artificial reference clean speech signal parameter. The speech degradation was estimated by finding the average distance between the parameters of the test and reference parameters.

One of the first non-intrusive signal based model was proposed by Liang and Kubichek , where perception based speaker independent speech parameters such as perceptual linear prediction coefficients (PLP) and perceptually weighted Bark spectrum are utilized for nonintrusive speech quality evaluation. The artificial reference parameters corresponding to high quality speech are derived from a variety of clean source speech material. They compared the degraded speech parameters to an artificial reference clean speech signal parameters that is appropriately selected from an optimally clustered codebook. The speech degradation was estimated by finding the average distance between the parameters of the test and reference parameter sets.

Some recent algorithms are based on GMM of features derived from perceptually motivated spectral envelope representations are modeled and used as an artificial reference to compute the distance from degraded speech[3]. The inconsistency between the artificial reference and degraded speech is mapped to the objective speech quality score using multivariate adaptive regression spline functions. An artificial reference model is created by using high quality undistorted speech based on PLP features and GMM clustering. Both narrowband and wideband speech quality rating depending on the inconsistency between the PLP features of GMM reference model and degraded speech are mapped using the support vector regression (SVR) method[4]. The perceptual spectral features are extracted from the degraded speech, which include temporal variations of speech as well as spectral statistical characteristics in the perceptual domain.

III. Proposed Methodology

The process starts with the data collection of audio signals. The dataset of audio signals was collected and preprocessed as per the requirement. After preprocessing of data, the dataset was split into training and testing data set. The model was trained on processed data and tested using signals. The project proposes Non-Intrusive Speech Quality assessment algorithm using support vector machine (SVM). This algorithm uses Mel-frequency cepstral coefficients (MFCCs) as speech features for speech quality assessment. The dataset is created which consists of various audio signals along with their corresponding MFCC features and subjective mean opinion scores (MOS). Support Vector Machine (SVM) algorithm is used for training of the database.This trained model gives speech quality score as output. Fig.1 shows generalized block diagram of our project.

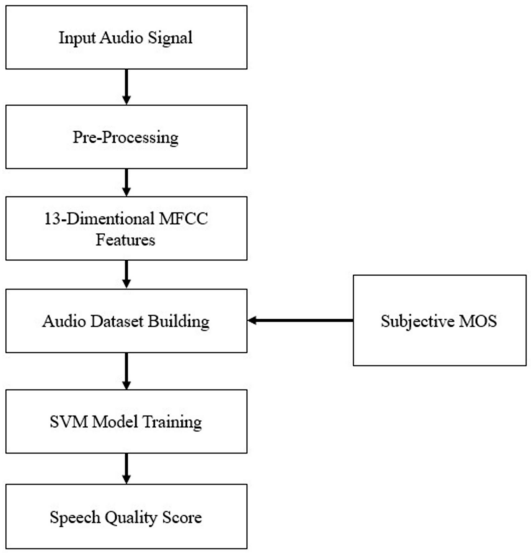

Fig. 1 : Generalized Block Diagram

A. Data Collection

Non - intrusive speech quality assessment model is developed using the machine learning algorithm. For training this algorithm data set is required. For the model development MFCC features are considered as key features. MFCC features of standard audio signals are extracted and the same signal's subjective MOS are computed. This two features are mapped together to build data set. For data set building Subjective speech quality assessment

experiment is performed. Live recorder has been included to record the audio signals. The dataset is divided into two parts such as training data and testing data. This training data is feed to the SVM model and training is performed. Accuracy of the model is then tested using testing data.

A. Machine Learning Model

1. SVM model

Support Vector Machine is supervised machine learning algorithm. Majority applications which require clarification of data points precisely then the SVM algorithm works efficiently. SVM separate out data points through a boundary plane into different classes of objects. Non - Intrusive speech quality assessment model is developed using Support Vector Machine algorithm. Audio classification problems are efficiently solved by SVM algorithm. SVM algorithm classifies data depending on boundary between data points. The plane which classifies the data point is called as hyperplane. The data set created using subjective MOS and MFCC features is trained using SVM machine learning algorithm. MOS scores from data set are considered as dependent variable and MFCC features are considered as independent variable. The dataset is divided into two parts such as training data and testing data. This training data is feed to the SVM model and training is performed. Accuracy of the model is then tested using testing data.

2. Feature Extraction

MFCC is a feature extraction method that is used to extract the features of audio signal. Here we have extracted 39 dimensional MFCC features out of which only 13 dimensional features are selected as they contain vocal tract contents which are required. The SVM model takes MFCC features as input. Model analyzes input audio file and compare it with data set present in the trained machine learning model and generates speech quality score. To find out the speech quality score user will have to record his/her audio through microphone. Then the MFCC features of this recorded audio signal are computed this features then provided to the SVM classifier. This model will predict the speech quality score depending upon its features. Fig.2 shows the System Architecture of our proposed model.

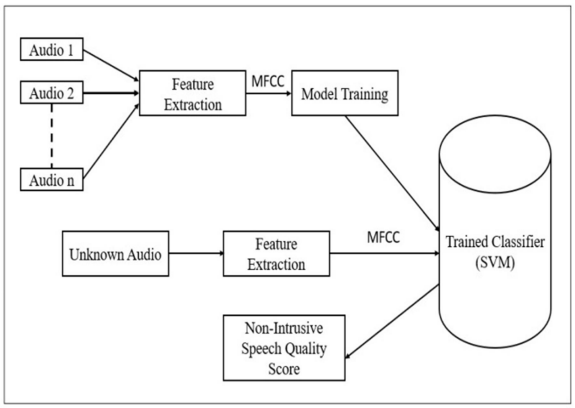

Fig.2 : System Architecture

IV. Results and Discussion

At present there is no open dataset for audio signals, we have created dataset through online resources. The support vector machine model is trained using dataset that is created using subjective listening experiment. 80% data of the total dataset is used for training and 20% data is used for testing. Testing data is passed to trained model and speech quality score is obtained as an output. The features computed for standard audio signal are shown in fig.3 which are fed to the model. The model had predicted score 2.7 which imply that signal is having lots of disturbance.

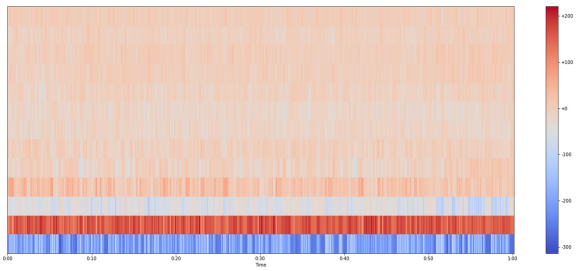

Fig.3 : MFCC Features for Standard signal

The same process is followed for the recorded audio whose mfcc features are shown in fig.4.The model had predicted speech quality score for this signal, score was found to be 3.28 which states that signal is having noise in it but not as much compared to the standard signal.

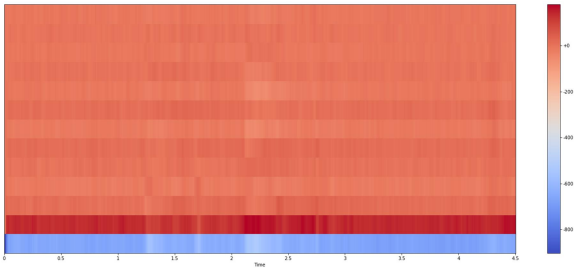

Fig.4 : MFCC Features for Recorded signal

Proposed Non-intrusive Speech Quality Assessment Algorithm predicted speech quality score with 70.66% accuracy. The predicted speech quality score lies between 0.5 and 4.5. If score is close to 0.5 then the quality of speech is poor which means signal is highly distorted whereas if score is close to 4.5 then the quality of speech is considered to be good which means signal has negligible distortions. The predicted score of our model is close to 3.5 which means the audio signal consists of certain distortions. Table.1 shows the score comparison of standard and recorded audio signals.

Audio Signal	Subjective MOS	Predicted Score
Standard Audio	2.5	2.7
Recorded Audio	2.9	3.28

Table.1 : Score comparison of standard and recorded audio signals

V. Conclusion

The recent improved Non-Intrusive Speech quality assessment models have shown better results in various telecommunication applications. . Speech quality differs from person to person as everyone may have different opinions about the same speech or sound. Hence the speech quality assessment is particularly subjective. The quality of speech may get affected due to various distortions and background noise. Thus, to avoid such factors we have developed a model that can maintain the quality of speech. The upsides of the proposed system are that it can really predict speech quality without reference to the clean signal and is successful in predicting quality of speech in terms of speech in high connection with subjective scores. The experimental results have exhibited its improvement over the most relevant existing measurement.

VI. Future Scope

MFCC feature extraction method using SVM is great algorithm as compared to previous ones but still now there are more improved algorithms which can give better results which will be improved in future as well. Developing a machine model in such a way that it can understand the overall quality of speech signal and evaluate the environment in which audio was captured without using the clean reference signal that is in non-intrusive manner is a challenging task which has many real world applications. However Neural Networks can be applied to the speech quality assessment to achieve good performance. So further this project can be improved by using neural network based algorithms to improve the quality of speech.

VII. References

[1] R.K. Dubey, A. Kumar, Non-intrusive objective speech quality evaluation using features at multiple time scales, in: Proceedings of the Acoustics, New Delhi, 2013, pp. 1040–1046.

[2] Cauchi, B., Siedenburg, K., Santos, J.F. et al. (3 more authors) (2019) Non-intrusive speech quality prediction using modulation energies and LSTM-network. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27 (7). pp. 1151-1163. ISSN 2329-9290

[3] R.K. Dubey, A. Kumar, Non-intrusive speech quality assessment using several combinations of auditory features, Int. J. Speech Technol. 16 (1) (2013) 89–101

[4] Dushyant Sharma, Yu Wang, Patrick A. Naylor, and Mike Brookes, Data driven Non-Intrusive measure of speech quality & intelligibility, Speech communication (volume 80) by European association for Signal Processing (EURASIP), 2016

[5] AndersonR. Avila,Hannes Gamper, Chandan Reddy , Ross Cutler , Ivan Tashev , Johannes Gehrke, NON-INTRUSIVE SPEECH QUALITY ASSESSMENT USING NEURAL NETWORKS,

Institute National de la Recherche Scientifique, Montreal, QC, Canada 2Microsoft Research Labs, Redmond, WA, USA[9], 2019

[6] Xuan Dong and Donald S. Williamson, A Classification-aided Framework For Non-intrusive Speech Quality Assessment, 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

[7] Haemin Yang, Kyungguen Byun and Hong-Goo Kang, "Parametric-based non-intrusive speech quality assessment by deep neural network", IEEE International Conference on Digital Signal Processing, 2016

[8] Z. Wanli and L. Guoxin, "The research of feature extraction based on mfcc for speaker recognition," 3rd International Conference on Computer Science and Network Technology, 2013.

[9] R. K. Dubey and A. Kumar, "Comparison of subjective and objective speech quality assessment for different degradation / noise conditions," in 2015 International Conference on Signal Processing and Communication (ICSC), 2015, pp. 261–266.

[10] V. P. Yousef Ettomi Ali and P. Doyle, "Disordered speech quality estimation using linear prediction," IEEE, 2017.