International Journal of Speech Technology (2022) 25:947–965
https://doi.org/10.1007/s10772-022-10011-y

Non-intrusive speech quality assessment using context-aware neural networks

Rahul Kumar Jaiswal[1] · **Rajesh Kumar Dubey**[2]

Received: 22 November 2021 / Accepted: 5 October 2022 / Published online: 23 October 2022

Abstract
To meet the human perceived quality of experience (QoE) while communicating over various Voice over Internet protocol (VoIP) applications, for example Google Meet, Microsoft Skype, Apple FaceTime, etc. a precise speech quality assessment metric is needed. The metric should be able to detect and segregate different types of noise degradations present in the surroundings before measuring and monitoring the quality of speech in real-time. Our research is motivated by the lack of clear evidence presenting speech quality metric that can firstly distinguish different types of noise degradations before providing speech quality prediction decision. To that end, this paper presents a novel non-intrusive speech quality assessment metric using context-aware neural networks in which the noise class (context) of the degraded or noisy speech signal is first identified using a classifier then deep neutral networks (DNNs) based speech quality metrics (SQMs) are trained and optimized for each noise class to obtain the noise class-specific (context-specific) optimized speech quality predictions (MOS scores). The noisy speech signals, that is, clean speech signals degraded by different types of background noises are taken from the NOIZEUS speech corpus. Results demonstrate that even in the presence of less number of speech samples available from the NOIZEUS speech corpus, the proposed metric outperforms in different contexts compared to the metric where the contexts are not classified before speech quality prediction.

Keywords Non-intrusive · Speech quality · Speech enhancement · Voice activity detector · Artificial neural network · Quality of experience

1 Introduction

On the face of global increment in the number of internet and mobile users around the world, the usages of Voice over Internet protocol (VoIP) applications, for example, Google Meet, Microsoft Skype, Apple FaceTime, etc. are growing with high pace. VoIP has become vital for the today's society including remote working, online communication like video conferencing, etc. To fulfill the user's expectations of better quality of experience (QoE) while using such VoIP applications, it is necessary to measure and monitor real-time speech quality. Various factors influence the QoE including system, network, content, and context of use (Falk et al., 2010). The service and system factors include the types of channels (mono or stereo), position of microphone, central processing unit (CPU) overload, etc. Jitter, packet loss and delay of the transmitted speech signal are included within the network factors. Content, that is, the characteristics of speech and voice may be affected by processing and can influence the QoE. The location of using a particular service comprises the contextual factors. For example, in the environments that are inherently noisy such as at the exhibition, restaurant, station or airport compared to the potential quietness at the home. *The contextual factors are primarily the centre of focus in the current research work.*

The traditional method to measure the quality of speech is absolute category rating (ACR; ITU, 1996), in which a number of subjects listen to the speech material played for them in a suitable environment and they give their quality

✉ Rahul Kumar Jaiswal
rahul.jaiswal@uia.no

Rajesh Kumar Dubey
rajesh.dubey@cuh.ac.in

[1] Department of Information and Communication Technology, Faculty of Engineering and Science, University of Agder, Grimstad, Norway

[2] Department of Electrical Engineering, School of Engineering and Technology, Central University of Haryana, Mahendergarh, India

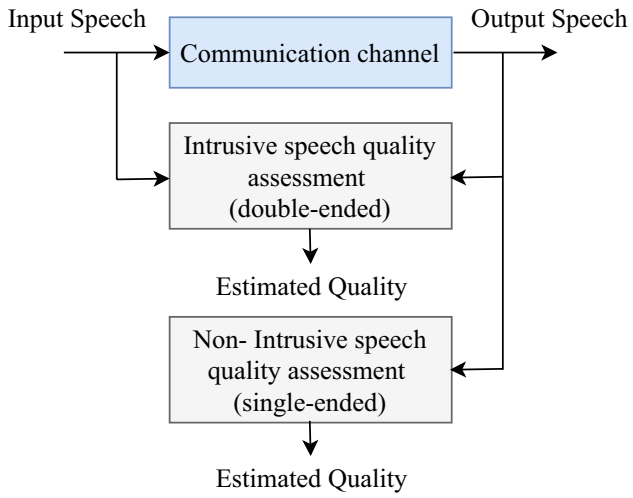

Fig. 1 Flow diagram of intrusive and non-Intrusive speech quality assessment metric

ratings. It is then averaged to obtain the mean opinion score (MOS). It is highly reliable and accurate method of obtaining speech quality ratings. But, it is time-consuming and impractical for real-time speech quality measuring and monitoring. Also, it is not possible to have a group of listeners at every node of speech communication networks for the speech quality subjective ratings (Holub et al., 2017). On the contrary, measuring and monitoring real-time speech quality using objective speech quality assessment metrics are more convenient, faster and practical.

Various objective speech quality assessment metrics are standardized by the International Telecommunication Union (ITU), for example, signal-based metrics which deploy received (degraded) speech signals for estimating the quality of speech. Two types of signal-based metrics exist (Möller et al., 2011): full-reference metric (also called "intrusive" or "double-ended" metric) and no-reference metric (also called "non-intrusive" or "single-ended" metric) as shown in Fig. 1.

Intrusive (reference-based) metrics usually calculate the distance between spectral representations of the transmitted reference (clean) signal and the received degraded signal. For example, the perceptual evaluation of speech quality (PESQ; Rix et al., 2001), the perceptual objective listening quality assessment (POLQA; ITU, 2011), the virtual speech quality objective listener (ViSQOL; Hines et al., 2015b) and its improved version ViSQOL v3 (Chinen et al., 2020) are some of the popular intrusive metrics. Since there is no access to the reference speech signal in real-time at receiver end and at different nodes of any telecommunication system, hence, the deployment of intrusive metrics for speech quality measuring and monitoring is not suitable for the telecommunication networks.

Non-intrusive (no-reference or reference-free) speech quality metrics (SQMs) are usually preferred for real-time measuring and monitoring of speech quality and the scenarios where the reference speech signal is not available. These metrics only deploy the degraded received speech signal to predict the quality of speech, and could be easily installed at the end point of VoIP channels for monitoring the quality of speech (Shome et al., 2019). Two no-reference objective SQMs are standardized for assessment of narrow-band speech signals (Möller et al., 2011). First, the ITU recommended P.563 (2004) and second, the American National Standard Institute (ANSI) standardized "auditory non-intrusive quality estimation plus (ANIQUE+; Kim & Tarraf, 2007)." ANIQUE+ is a perceptual model which simulates the functional roles of human auditory system and deploys improved modelling of quality estimation using a statistical learning paradigm (Kim & Tarraf, 2007). ANIQUE+ is only commercially available. P.563 is publicly available. However, both P.563 and ANIQUE+ metrics do not take into account the type/class of input speech signal while predicting the quality of speech.

Several non-standardized speech quality assessment metrics are developed in literature for non-intrusive objective speech quality evaluation. For example, the low complexity speech quality assessment metric (LCQA; (Bruhn et al., 2012), metric using multiple time-scale auditory features (Dubey & Kumar, 2017) and deep neural network (DNN) based speech quality assessment metrics (Avila et al., 2019; Catellier & Voran, 2020; Fu et al., 2018; Ooster et al., 2018; Soni & Patil, 2021; Wang et al., 2019). To monitor the quality of speech over a communication network, the LCQA algorithm deploys low complexity (execution time). In order to estimate the quality of speech, it maps the global statistical features, for example, mean, variance, skewness and kurtosis obtained from speech codecs using Gaussian mixture model (GMM). For each frame, the global features of speech signals are calculated from the speech-coding parameters (Grancharov et al., 2006). Reported results indicate that LCQA metric performs poorly in the presence of competing speaker (Jaiswal & Hines, 2018) type degradations. Moreover, the LCQA metric is restricted only to the parametric representation of the input speech signal without its perceptual transform. In Dubey and Kumar (2017), the author used a combination of different auditory features, such as Lyon's auditory features, Mel frequency cepstral coefficients (MFCC) and line spectral frequencies (LSF) and then trained a joint GMM for computing the quality of speech. However, the metric do not discriminate different types of degraded noisy speech signals before estimating the quality of speech as each type of speech/noise has different spectral characteristics and it can influence the estimation of speech quality. Some recent works on DNN-based speech quality assessment metrics include (Avila et al., 2019; Catellier & Voran,

2020; Fu et al., 2018; Ooster et al., 2018; Wang et al., 2019). For example, MFCC features are extracted and then used for training a DNN in order to predict the quality of speech in Avila et al. (2019), and a waveform-based convolutional neural network (CNN) is trained to predict the quality of speech in Catellier and Voran (2020). An output-based speech quality assessment metric incorporating autoencoder and support vector regression is implemented using NTT-AT Chinese corpus containing different types of noise degradations in Wang et al. (2019). Further, Soni and Patil (2021) predicts quality of speech from the speech signal by deploying deep autoencoder (DAE) and sub-band autoencoder (SBAE) features and then training an artificial neural network (ANN) on noisy speech samples obtained from the NOIZEUS speech corpus (Hu & Loizou, 2006). However, all these DNN-based metrics predict the quality of speech directly, that is, without identifying the type of noise class (context) of the input speech signal. Since each type of noise class has a separate behaviour and spectral characteristics, therefore, identifying the type of noise class (context) of speech signal and then switching to that noise class for predicting the noise class-specific (context-specific) speech quality could have a significant effect on the speech quality prediction accuracy.

Some metrics estimate the quality of speech using the network and the terminal parameters, known as, parametric metrics (Yang et al., 2016), for example, the E-Model (Bergstra & Middelburg, 2003). Network delay and packet loss are the parts of network parameters. Terminal parameters comprise jitter buffer overflow, coding distortions, jitter buffer delay, and echo cancellation. Using these parameters, impairments of the received speech signals are predicted and then rating factor is converted into mean opinion score[1] (MOS). However, the limitation of the E-Model includes its inability in representing the non-linear relationship between perceptual characteristics of speech signal and network planning parameters due to the dynamic change in the characteristics of speech signal. Moreover, the parametric metrics do not deploy the speech signal in quality predictions, therefore, unsuitable in predicting the quality of speech based on signal-noise characteristics (Möller et al., 2011). In order to meet the desired QoE of human while using VoIP applications, it is essential to have a no-reference signal-based speech quality assessment metric which should be aware of the type of noise class (context) of the input speech signal while measuring and monitoring real-time speech quality.

While digging thoroughly inside the literature, we find that there is no such non-intrusive signal-based speech quality assessment metric that can firstly identify the noise class (context) of the speech signal prior to predicting the quality of speech in real-time, therefore, we motivated to develop a noise class-awared QoE prediction metric for real-time prediction of speech quality and its efficient utilization in monitoring the quality of speech. Using artificial intelligence (AI) and machine learning (ML) techniques, smart decision making AI-based algorithms can be introduced into the mobile devices for improving the QoE and the performance gains of the end-user. To that end, our proposed non-intrusive noise class-awared speech quality prediction metric comprises of three main components. First, a noise class-classifier for classifying the noise class (context) of the input speech signal; second, a voice activity detector (VAD) for identifying the voiced segments inside the input speech signal; and third, noise class-specific or context-specific speech quality prediction metric (CSQM) for predicting noise class-specific speech quality of the input speech signal. For training the noise class-classifier and the CSQM excellently, it is desirable to have a large amount of noisy speech samples. However, due to the availability of small number of speech samples of different noise classes in the NOIZEUS speech corpus (Hu & Loizou, 2006), we have also addressed these challenges by developing a novel machine learning classifier algorithm for accurately classifying the noise class (context) of the speech signal and a collection of novel DNN-based CSQMs which are trained for each noise class to estimate the quality of speech.

The internet service providers can easily deploy our proposed speech quality prediction metric for measuring and monitoring the service performance quality continuously and can detect the impairments by identifying the noise class (context). This can assist in identifying the potential root causes and in installing the QoE-aware management actions to react and maintain the end-user QoE levels (Jahromi et al., 2018). The key contributions of the proposed work are as follows:

- Addressing speech quality monitoring problem by developing a novel noise class (context) aware SQM using three-step process: first step is to obtain noise-class (context) classifier; second step is to perform pre-processing using VAD; and third step is to develop noise-class (context) specific SQM. All the three steps have separate relevance.
- Developing a novel noise-class (context) classifier using a novel feature extraction technique to train the ML classifier.
- Developing a VAD algorithm to detect the presence of speech signal segments.
- Developing a group of DNN-based SQMs which are trained and optimized for a specific noise class, that is, noise-class (context) sensitive.
- Classifying the noise class (context) of the input noisy speech signal and then switching to a specific speech

[1] Mean opinion score (MOS) represents the rating of speech quality on a range from 1 (bad) to 5 (excellent) (see Table 3).

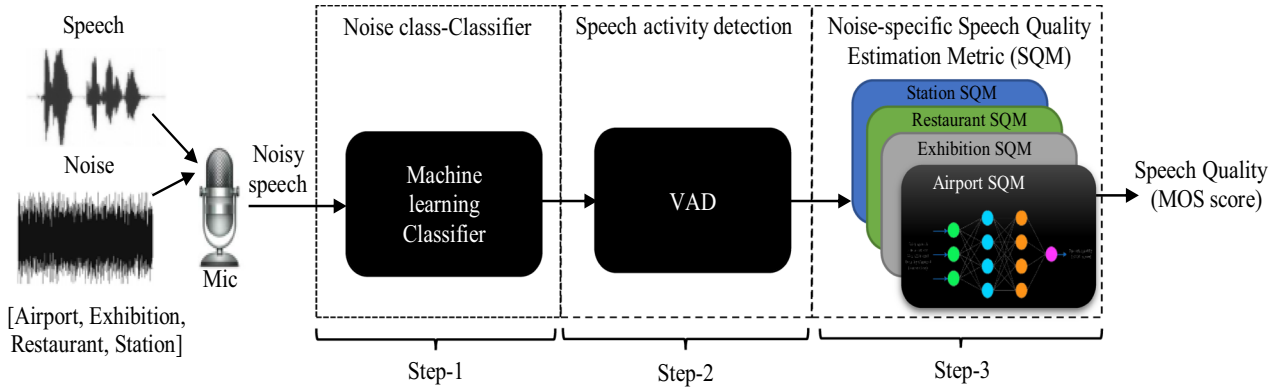

Fig. 2 Block diagram of the proposed speech quality prediction metric

quality model (SQM) results in a better speech quality prediction and it will allow the end-user to perceive a better QoE over VoIP applications.

– The proposed 3-step QoE framework shows improved performance and a significant advantage over the simple SQM where the speech quality is predicted directly without identifying the noise-class (context).

The remainder of this paper is structured as follows: Sect. 2 presents detailed explanation of the proposed SQM and its associated components. Section 3 describes the experimental dataset. The evaluation methodologies of each component and overall metric are described in Sect. 4. Section 5 presents the system design for executing the program and choice of different parameters and hyper-parameters. Section 6 presents and discusses the results of each component and overall metric before concluding remarks and future directions are presented in Sect. 7.

2 Proposed speech quality metric

Figure 2 shows each building block of the proposed noise class-awared speech quality prediction metric. It constitutes of mainly three sub-parts: (a) noise-class classifier to identify the noise class (context) of speech signal; (b) VAD to segregate the voiced and non-voiced segments from the speech signal; and (c) noise class-specific SQMs which are trained and optimized for each specific noise class using deep neural networks (DNNs), that is, context-aware DNN-based SQMs, in order to evaluate the quality of speech under that particular noisy scenario. Behind developing our proposed QoE metric, the hypothesis is: "With the prior knowledge of the noise class of the speech signal under test using a classifier, the test signal can be directed to the corresponding speech quality assessment metric which is trained and optimised for that specific noise degradation."

2.1 Noise-class classifier

The noise-class classifier is the first building block of the proposed metric. In developing robust classifier, we use the information that different speech enhancement algorithms (Hu & Loizou, 2006, 2007) have different associated noise estimation techniques (see Table 1), resulting in varying success while enhancing the noisy speech signals. The key is that with different levels of enhancements of noisy speech signals by different SE algorithms, their objective speech quality ratings (MOS scores) will also be different. With this prior information, each input noisy speech signal is processed via 12 standard speech enhancement algorithms (Hu & Loizou, 2006, 2007) as outlined in Table 1. These 12 processed speech signals along with the one original unprocessed noisy speech signal results in 13 different variations for each input signal. These 13 signal variants (12 processed and 1 unprocessed) are then injected to the P.563 metric (ITU, 2004) (P.563 is discussed next) for acquiring 13 different speech quality ratings, called "MOS score" (see Table 3). Further, these MOS scores are integrated in order to deploy it as input feature vector for training the classifier and then identifying the noise class. This approach of feature extraction from the noisy speech samples for training a classifier in order to detect the noise class is the only one of its kind in the literature. The block diagram of the noise-class classifier illustrating feature extraction technique is shown in Fig. 3. In the following sub-sections, two sub-parts of the noise-class classifier (speech enhancement algorithms and objective SQM) and used ML classifier are discussed.

2.1.1 Speech enhancement

For improving the speech signals degraded by different types of noises, speech enhancement (SE) algorithms are deployed (Das et al., 2020; Moore et al., 2017; Saleem & Khattak, 2019). The performance of SE algorithms depends on the

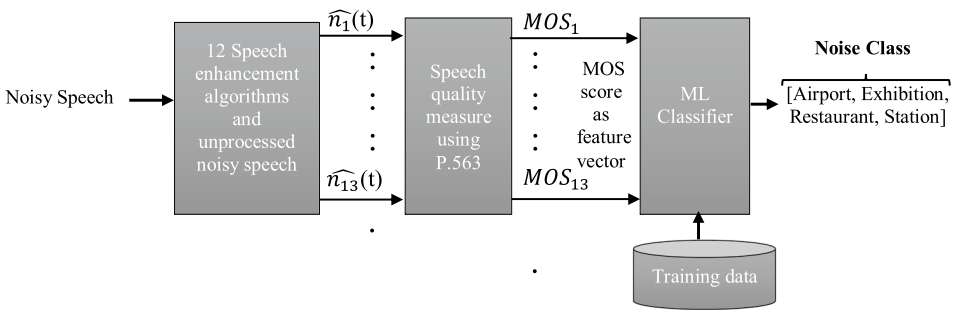

Fig. 3 Block diagram of noise-class classifier illustrating feature extraction technique to train the classifier (Jaiswal & Hines, 2020)

Table 1 Class of speech enhancement algorithms with associated noise estimation techniques (Hu & Loizou, 2006, 2007)

Class of speech enhancement algorithms	Number	Noise estimation techniques
Adaptive filtering	3	Wiener filtering (Hu & Loizou, 2004) and A priori signal to noise estimate (Scalart et al., 1996)
Spectral subtraction (Kamath & Loizou, 2002)	2	Adaptive gain averaging and reduced delay convolution (Gustafsson et al., 2001)
Statistical model-based (Ephraim, 1992)	5	Minimum mean square error (Ephraim & Malah, 1984) and log-MMSE (Cohen, 2002)
Signal-subspace (Hu & Loizou, 2003)	2	Karhunen–Loeve transform (KLT) (Mittal & Phamdo, 2000)

characteristics of surrounding noise and noise estimation algorithms associated with it. The 12 standard SE algorithms presented in Hu and Loizou (2006, 2007) consist of four classes. Each class of SE algorithm in Hu and Loizou (2006, 2007) involves a separate noise estimation technique resulting in varying success in enhancing the degraded speech. The different classes of SE algorithms with associated noise estimation techniques are presented in Table 1.

2.1.2 ITU-T P.563 metric

ITU standardized P.563 (2004) is a no-reference speech quality assessment metric designed for estimating quality of active speech in narrow-band speech signals. Three main principles (Malfait et al., 2006) define the P.563 metric. First, physical model of vocal tract; second, reconstruction of the reference signal to evaluate unmasked distortions; and third, focus on specific distortions e.g., temporal clipping, robotization, and noise. P.563 involves several steps in predicting the quality of speech which includes pre-processing, detection of dominant distortion class, and perceptual mapping. The pre-processing includes reverse filtering, adjustment of speech level, identification of speech portions and calculation of speech and noise levels using a VAD (ITU, 2004). Distortion class includes unnaturalness of speech, robotic voice, beeps, background noise, signal-to-noise ratio (SNR), mutes, interruptions, extracted from the voiced parts of speech signals. Finally, a dominant distortion class is detected and mapped to a single mean opinion score denoted as "mean opinion score of objective listening quality (MOS-LQO)" as presented in Table 3. Table 2 presents the requirements of speech signals when evaluating using P.563.

Table 2 Requirements of speech signals in P.563 (ITU, 2004)

Sampling frequency	8000 Hz
Amplitude resolution	16 Bit linear PCM
Minimum active speech in sample	3.0 s
Maximum signal length	20.0 s
Minimum speech activity ratio	25%
Maximum speech activity ratio	75%
Range of active speech level	−36.0 to −16.0 dBov

Table 3 Speech quality rating (MOS Scale)

Speech quality	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

2.1.3 Classifier used

A major challenge in our data-driven approach is to develop an accurate and robust classifier with small number of speech samples that should have the ability to detect the noise class of input speech signal correctly. To that end, different ML classifiers (Reddy et al., 2021; Shami & Verhelst, 2007; Singh & Singh, 2021), for example, XGBoost,

decision tree, random forest, logistic regression, K-nearest neighbor, support vector machine, and Naive Bayes are investigated as per the classification approach in Jaiswal and Hines (2020) where it has been performed for 8 noise classes instead of 4 noise classes. The simulation results for 4 noise classes (see Tables 6, 7) report that XGBoost classifier has high accuracy in classifying the noise class of speech signal. Therefore, we deploy the XGBoost classifier.

The extreme gradient boosting (also called XGBoost) is ensemble of classification and regression trees (CART; Chen & Guestrin, 2016). In XGBoost, the trees are optimized using gradient boosting technique (Friedman, 2001) which minimises the loss function of the model by adding weak learners using gradient descent. Let the output of a tree be:

$$f(x) = w_q(x_i), \tag{1}$$

where x is the input vector and w_q is the score of the corresponding leaf q. Then, the output of an ensemble of K trees will be (Dimitrakopoulos et al., 2018):

$$y_i = \sum_{k=1}^{K} f_k(x_i). \tag{2}$$

The XGBoost algorithm tries to minimize the following objective function J at step t (Dimitrakopoulos et al., 2018):

$$J(t) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i), \tag{3}$$

where the first term contains the train loss function L (e.g. mean square error for regression and binary cross entropy for classification) between the actual class y and the predicted output $\hat{y}$ for n samples and the second term is the regularization term which helps in controlling the complexity of the model and avoiding overfitting. With T being the number of leaves, γ being the pseudo regularization hyper-parameter, and λ being the $L2$ norm for leaf weights, the complexity $\Omega(f)$ is defined as (Dimitrakopoulos et al., 2018):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {w_j}^2. \tag{4}$$

Our anticipation is that the relationship between the unprocessed speech quality estimates and the enhanced speech quality estimates would be learnt by the classifier in correctly classifying the noise class of input speech signal.

2.2 Voice activity detection

The VAD is the second building block of the proposed metric. Based on the speech features, VAD identifies the active voiced segments in the input speech signal (Fukuda et al., 2018) as shown in Fig. 4. The speech quality and

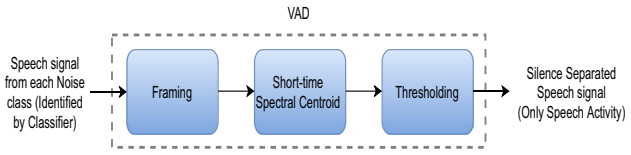

Fig. 4 Block diagram of used WS VAD illustrating feature extraction technique to segregate silences from the speech signal

intelligibility are based on the voiced segments of the input speech signals only. It works as a pre-processing unit for processing the input speech signal in order to segregate silences prior to injecting it to the speech quality estimation component. It has been noticed in our previous study (Jaiswal, 2022) that the weighted spectral centroid (WS) VAD performs excellent as compared to other VADs in detecting and separating the voiced segments from the noisy speech signal.

The weighted spectral centroid VAD extracts spectral centroid (SC) feature from the speech signal. Spectral centroid is the centre of mass of the spectrum and its high value corresponds to the "brightness" of the sound. In order to develop the WS VAD, the speech samples are divided into overlapping frames of size 25 ms with 10 ms shift and then short-time spectral centroid is calculated for each frame as (Jaiswal, 2022):

$$SC_i = \frac{\sum_{k=1}^{N}(k+1)X_i(k)}{\sum_{k=1}^{N} X_i(k)}, \tag{5}$$

where $X_i(k)$, $k = 1, 2, \ldots, N$ is the discrete Fourier transform (DFT) coefficients of the ith short-time frame of signal X having frame-length N. After calculating short-time SC of each frame, a smoothing filter, that is, median filter[2] (Deligiannidis & Arabnia, 2014) is applied throughout the feature sequence. Then, histogram is computed and its local maximas are detected. Thereafter, threshold T is measured dynamically using the weighted average between the first and the second local maxima as, $T = \frac{W_1 M_1 + W_2 M_2}{W_1 + W_2}$, where M_1 and M_2 are the first and the second local maxima, respectively. W_1 and W_2 are the user-defined weights, and are set as $W_1 = 5$ and $W_2 = 1$. The measured threshold is applied to each frame for extracting the voiced segments of the speech signal.

It is important to mention here that pre-processing stage in P.563 (ITU, 2004) includes an internal VAD which is based on the adaptive power threshold, using an iterative approach as shown in Fig. 5. However, this internal VAD only computes a range of features and the active and inactive portions of speech activity but does not pre-process to

[2] A median filter is a non-linear filtering technique used to remove noise from the signal (Deligiannidis & Arabnia, 2014).

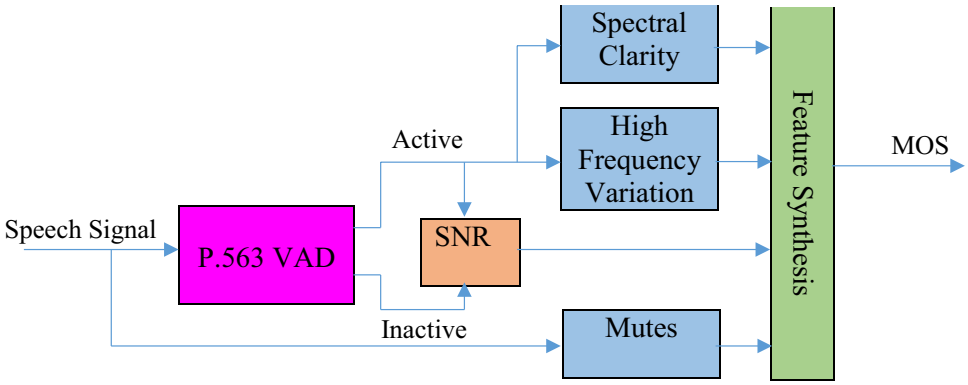

Fig. 5 VAD in P.563 (Jaiswal & Hines, 2018)

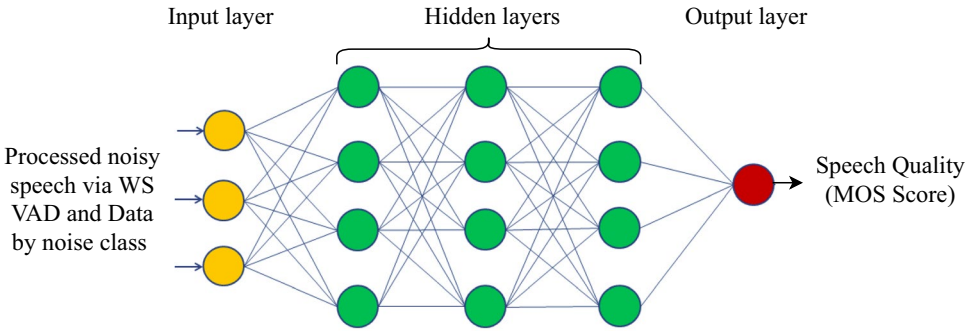

Fig. 6 A simple three layered feed-forward DNN to develop noise-class (context) specific speech quality metric (SQM)

remove silences from the speech sample which can be seen in full details of P.563 specifications (ITU, 2004). Also, most of the VAD algorithms fail when the level of background noise increases (Ramirez et al., 2007). For example, it has been seen in Hines et al. (2015a) that the P.563 metric performs poorly in estimating speech quality on speech samples containing silences. Therefore, we deploy WS VAD (in Step-2) to pre-process the speech samples prior to developing a collection of noise class-specific (context-specific) SQMs in Step-3.

2.3 Speech quality estimation

A collection of DNN-based noise class-specific SQMs, trained and optimised for a particular noise class (context) are the third building block of the proposed metric. The fully-connected DNNs and feature selection techniques, for example, lasso and ridge are deployed for its development.

2.3.1 Fully-connected DNNs

A DNN is an artificial neural networks (ANNs) consisting of L layers, including one input layer, $L-2$ hidden layers, and one output layer as shown in Fig. 6. The output of the DNN is a cascade of non-linear transformation of the input. Consider a feed-forward DNN with L layers, labelled $l = 1, \ldots, L$ and each having corresponding dimension q_l. The layer l is defined by the linear operation $W_l \in \mathbb{R}^{q_{l-1} \times q_l}$ followed by a non-linear activation function $\sigma_l : R^{q_l} \rightarrow R^{q_l}$. Layer l receives input from the $l-1$ layer denoted as, $w_{l-1} \in R^{q_{l-1}}$, the resulting output of the layer l, $w_l \in R^{q_l}$, is then computed as $w_l := \sigma_l(W_l w_{l-1})$, where $\sigma_l(\cdot)$ is the point-wise activation function. The final output of DNN, w_L, is then related to the input w_0 by propagating through various layers of the DNN as $w_L = \sigma_L(W_L(\sigma_{L-1}(W_{L-1}(\ldots(\sigma_1(W_1 w_0))))))$, where the DNN learns layer-wise weights $w_1, w_2, \ldots, w_L$ (Eisen et al., 2018).

For developing a collection of noise-class (context) specific speech quality metrics (SQMs), noisy speech signals processed through WS VAD (that is, silence separated speech signals) and noise-class specific training data are fed to the input of DNN as shown in Fig. 6. It is important to note here that zero padding (Engelberg, 2008) is used at the end of the processed speech samples in order to make the length of each processed speech samples same before feeding to the DNN. The output of the DNN is the speech quality prediction (MOS score). The activation function of hidden layers σ_l include rectified linear unit (ReLU), defined as $\sigma_l(x) = 0$ for $x < 0$ and x for $x > 0$, and linear, defined as $\sigma_l(x) = cx$ where c is a constant. It varies for each noise class. The output layer σ_o includes a linear activation function. The noise-class classifier acts as a filter and firstly classifies the noise class of input speech signal and then activates the corresponding trained DNN in order to obtain the predicted speech quality (MOS score).

2.3.2 Lasso feature selection

Least absolution shrinkage and selection operator (Lasso; Li et al., 2017) is a high-dimensional feature selection technique used to separate the relevant features from the irrelevant ones and helps in reducing the complexity of the model. It uses cross-validation to adjust the tuning parameters along all aspects of the model and to prevent overfitting. It is a linear regression method that minimizes the usual square error loss plus $L1$ penalty on the regression coefficients. The minimization of the overall loss function is given as (Jain et al., 2014):

$$\hat{\beta} \leftarrow \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{n=1}^{N} \left(y_n - b - \sum_{d=1}^{D} \beta_d x_{dn} \right)^2 \right\} + \lambda ||\beta||_1, \quad (6)$$

where $\{x_n, y_n\}_{n=1}^{N}$ is the training data, b is the intercept, and the Lagrange multiplier λ balances the trade-off between the squared error loss and the $L1$ penalty $||\beta||_1$ on the regression coefficients. The $L1$ norm imposes a competition between the regression coefficients, resulting in shrinking some of them to 0, and therefore producing a sparse model that explains the data with as little features as possible. The λ is typically inferred from cross validation. This optimization is performed using the 5-fold cross validation.

2.3.3 Ridge feature selection

Ridge (Chowdhury et al., 2018) is also one of the high-dimensional feature selection techniques used to obtain the relevant features for training the DNN models smoothly with reduced complexity. It is a variant of the regularized least squares problems where the choice of the penalty function is the squared $L2$ norm. With $A \in \mathbb{R}^{nxd}$ being the design matrix, $b \in \mathbb{R}^n$ being the response vector, and $\lambda > 0$ being the regularization parameter, the ridge coefficients minimize penalized residual sum of squares, given as (Chowdhury et al., 2018):

$$Z = \min_{x} \left\{ ||Ax - b||_2^2 + \lambda ||x||_2^2 \right\}. \quad (7)$$

Solving standard least-squares problems without regularization may provide a good fit with the training data but may not generalize well with the test data. Therefore, to address this problem, ridge regression waives off the requirements of unbiased estimator. At the cost of introducing bias, ridge regression reduces the variance and thus can reduce the overall mean square error (MSE). The tuning parameters are adjusted properly to all aspects of the model using efficient leave-one-out (Meijer & Goeman, 2013) cross-validation technique, which helps in preventing overfitting as well.

3 Experimental dataset

In daily life, noise appears in different shapes and forms. Each type of noise has a separate behaviour and spectral characteristics. For example, fan noise coming from the personal computer is stationary whereas restaurant noise, where multiple people speak in the background, is non-stationary due to its constantly changing spectral characteristics. Therefore, improving the speech quality of the speech signal degraded by non-stationary noise is more difficult than stationary noise.

Different datasets have different applicabilities. For example, ITU-T P.Supplement-23 database (1998) contains the coded version of speech utterances used in the ITU-T 8 kbps codec characterization tests (Dubey & Kumar, 2013). The test performed three experiments. The Experiment-1 examines G.729 codec with coded speech samples, thus, not useful for our study. The Experiment-2 investigates the effect of background noise for transmission quality and the method of assessment used here is comparison category rating (CCR) not ACR, thus, not suitable for our study. Experiment-3 investigates effects of channel degradations using coded speech samples, thus, also not useful for our study. Hence, this dataset is not suitable for testing our proposed QoE metric due to the unavailability of different types of noise degradations present in speech signals under noisy environments. Furthermore, there is no large amount of open-source real-world noisy speech dataset with listener quality rating or subjective quality rating that can help in the assessment of our proposed metric. Obtaining subjective quality rating of large noisy dataset is a cumbersome work.

Since the speech signal is degraded with different types of environmental noises, therefore, a noisy speech dataset is needed to investigate the performance of the proposed QoE metric. NOIZEUS (Hu & Loizou, 2006) is an open-source noisy speech dataset containing different types of noise degradations. It has 30 phonetically-balanced IEEE English sentences, pronounced by three male and three female speakers. Each sentences are degraded with four types of commonly occurring real-world noises, for example, airport, exhibition, restaurant and station at two SNRs, that is, 5 dB and 10 dB. The noises are taken from the AURORA database (Hirsch & Pearce, 2000). Each sentences are down-sampled from 25 to 8 kHz, that is, narrow-band speech samples. All the speech samples are saved in .wav format (16 bit PCM, mono) and the average duration of each utterance is three second.

4 Evaluation of the proposed speech quality metric

In this section, we present the evaluation methodology used for the evaluation of the proposed metric and its associated building blocks.

4.1 Evaluation of the noise-class classifier

To evaluate the noise-class classifier, we take 30 noisy speech samples available from the NOIZEUS speech corpus having four real-world noise classes, that is, airport, exhibition, restaurant and station at two SNRs, that is, 5 dB and 10 dB. Thus, we have 60 (30 samples × 2 SNRs) noisy speech samples for each noise class, resulting in total 240 (30 samples × 4 noise classes × 2 SNRs) noisy speech samples. For extracting the features in order to train noise-class classifier, each noisy speech sample is processed through 12 standard speech enhancement algorithms (see Table 1), resulting in 12 processed speech samples along with one unprocessed original noisy speech sample, that is, total 13 (12 + 1) variants of speech samples. Next, these 13 variants of speech samples, that is, $\hat{n_1}(t), \hat{n_2}(t), \ldots, \hat{n_{13}}(t)$ are injected to the P.563 metric (see Fig. 3) to obtain 13 different objective speech quality predictions, that is, $\text{MOS}_1, \text{MOS}_2, \ldots, \text{MOS}_{13}$. We, then, integrate these 13 MOS scores and deploy them as the input feature vector to train the classifier for classifying the noise class (context) of the given test speech sample.

We have only 60 noisy speech samples in each noise class,[3] which is very small for training a data-driven classifier accurately. There are four noise classes. Training with a multi-class ML classifier results in a poor classification accuracy of only 38%. Therefore, we perform one-vs-all approach of multi-class classification, that is, binary classification with imbalanced datasets for each noise class as per the strategy shown in Fig. 7. We assign the first noise class (e.g., Airport) as "class 0" and the remaining three noise classes as "class 1", and label it as "Airport". Similarly, we assign the second noise class as "class 0" and the remaining three noise classes as "class 1", and label it as "Exhibition". We follow the same strategy for the remaining noise classes. This makes the binary class samples imbalanced. Therefore, for balancing both noise classes, we reduce the size of majority noise class (class 1) equals to the size of minority noise class (class 0) using the under-sampling technique (Drummond & Holte, 2003; Fernández et al., 2018). Once both noise classes are balanced, we divide the data of each noise class into 80:20 ratio for training and testing the classifier.

To evaluate the performance of classifier, F-score (FS) or test accuracy and geometric mean (G-mean) or balanced accuracy are investigated. F-score (Belarouci & Chikh, 2017) is the weighted harmonic mean of precision (PR) and recall (RC). G-mean (Belarouci & Chikh, 2017) is the geometric average of the classification precision of the minority

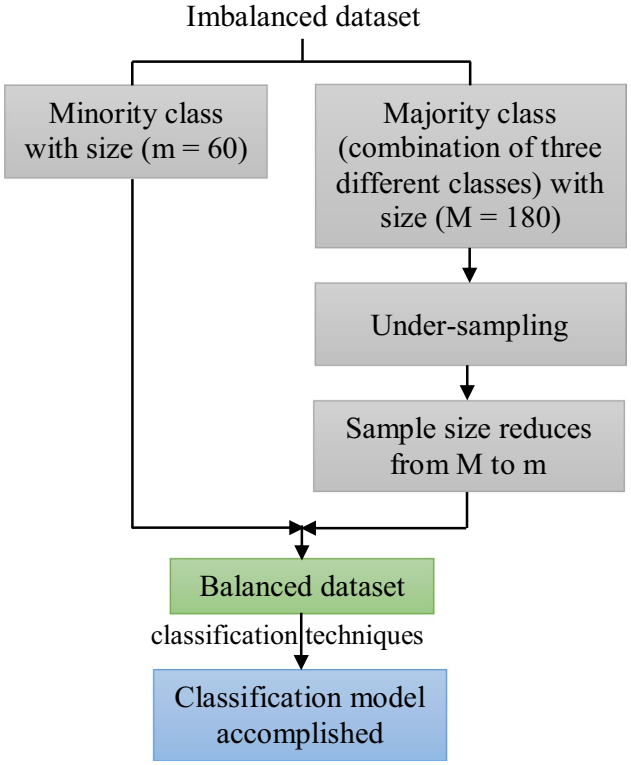

Fig. 7 Flow diagram of binary classification with imbalanced dataset

and the majority class. It evaluates the model's ability to correctly classify the minority and the majority class. With TP, TN, FP, and FN being the true positive, true negative, false positive and false negative, respectively, the F-score and G-mean are given as (Jaiswal & Hines, 2020):

$$PR = \frac{TP}{TP + FP}, \quad RC = \frac{TP}{TP + FN}, \quad FS = \frac{2(PR \times RC)}{(PR + RC)}, \tag{8}$$

$$G\text{-mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}. \tag{9}$$

4.2 Evaluation of the VAD

To evaluate the performance of VAD, we obtain the frame-wise binary mask[4] of noisy speech samples using the WS VAD and compare it to the ground truth (GT) VAD mask[5]

[3] A noise class refers to noisy speech samples obtained by combining noisy samples at two SNRs of 5 dB and 10 dB e.g., Airport.

[4] Binary mask is binary decision taken by a VAD. If the measured value exceeds the threshold then VAD = 1, that is, voiced segments, else, VAD = 0, that is, noise/silence.

[5] Ground truth (GT) VAD mask is the ideal binary mask which is computed as silence, if the frame's sample value = 0; and voiced segments, otherwise; for the reference (clean) speech samples.

to obtain the TP, TN, FP and FN for calculating precision (PR), recall (RC) and F-score (FS) (Jaiswal & Hines, 2020). F-score measures the test accuracy and it's maximum value is 1 (PR = RC = 1) which signifies that the voice activity decision of the VAD algorithm is equal to its reference transcription.

4.3 Evaluation of the noise-class specific speech quality metric

The availability of sufficiently large learning datasets is the key to the success of DNN models in replicating different optimization-based speech quality solution. However, it is not possible to obtain large amount of learning datasets openly in order to satisfy the data hungry nature of the DNN models because most of the real-data is publicly unavailable and/or costly to obtain. Alternatively, one can generate realistic data using advanced deep learning techniques, such as, generative adversarial networks[6] (GANs; Goodfellow et al., 2014). However, that is beyond the scope of this research.

Further, the processed speech samples of each noise class via WS VAD and the training data of the corresponding noise class (see Fig. 3) are injected as the input to the DNNs (see Fig. 6). The subjective speech quality predictions (MOS-LQS) of each noise class obtained from Dubey and Kumar (2015) are the output to the DNNs. A collection of DNNs (see Fig. 6) are trained and optimised for each noise class to obtain noise-class specific speech quality prediction metrics. The motivation for exploiting DNNs is primarily due to its universal approximation capability (Sun et al., 2017), supplemented by the fact that trained DNN models are computationally simple (Ye et al., 2017) to execute.

Next, the lasso and ridge feature selection techniques are used to extract the most appropriate feature for training the DNN of each noise class. The lasso feature selection technique is used for station class and the ridge feature selection technique is used for airport, exhibition and restaurant noise class. The layers of the DNNs are densely connected. Various combinations of hidden layers and weights are experimented for achieving the best DNN model of each noise class that could be trained in reasonable time. For the performance evaluation of the DNNs, the training and testing mean square error (MSE) of each noise class are computed. MSE is defined as the average of the square of the differences between the actual (subjective) and the predicted (objective) quality scores. With n being the number of speech samples, $y_1, y_2, \ldots, y_n$ and $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ being the actual (subjective) and the predicted (objective) quality scores, respectively, MSE is given as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{10}$$

4.4 Evaluation of the overall proposed metric

The Pearson correlation ρ_p, Spearman correlation ρ_s and root mean square error (RMSE) between the objective quality scores (MOS-LQO) obtained by the proposed metric for each test sample of each noise class and their corresponding subjective listener quality scores (MOS-LQS) are used for the performance evaluation of the overall proposed metric. With μ_y and $\mu_{\hat{y}}$ being the mean of subjective and predicted quality scores respectively, and d_i being the difference between ranks of the subjective and the predicted quality scores, these measures are given as (Falk & Chan, 2006; Sharma et al., 2016):

$$\rho_p = \frac{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})^2(y_i - \mu_y)^2}}, \tag{11}$$

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{12}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{13}$$

5 System design

This section describes the simulation platform used and different choices of parameters and hyper-parameters for training the classifier and the DNN models to obtain the proposed SQM. MATLAB 2018a is used to implement the VAD algorithm. The standard P.563 metric is programmed in "C" language. The classifier and the DNN-based SQMs are implemented in Python 3.7.3 having TensorFlow 2.2.0 on Windows 10 laptop with Intel Core i5 8th generation processor, Intel UHD Graphics 620, and 16 GB of memory.

5.1 Choice of parameters and hyper-parameters

To obtain the noise-class classifier which can accurately classify the noise class of the noisy speech signal and to develop a collection of robust DNN-based SQMs trained for each noise class, different parameters and hyper-parameters are experimented in our system design. Different search algorithms, for example, grid search, and random search

[6] Generative adversarial networks (GANs) is a new class of generative methods for data distribution learning where the objective is to learn a model that can generate samples close to the target distribution (Ye et al., 2018).

Table 4 Parameters and hyper-parameters of XGBoost classifier for each noise class

Name of parameters	Airport noise class	Exhibition noise class	Restaurant noise class	Station noise class
Number of estimators	9	100	50	11
Maximum depth	6	6	3	3
Learning rate	3	1	0.4	2.7
Subsample	0.918803760	0.814572864	0.736947389	0.456651330
Minimum child weight	0.986197239	0.667779632	0.658131626	0.999199839

(Bergstra & Bengio, 2012) are also performed to obtain the most appropriate parameters.

For the tested XGBoost classifier, the choices of parameters and hyper-parameters are presented in Table 4.

To obtain the robust DNN-based SQMs trained and optimized for each noise class, that is, context-specific speech quality metrics, different number of hidden layers with variable number of neurons are experimented. Rectified linear unit (ReLU) and linear activation function are used for the hidden layers depending on each noise class. Linear activation function is used for the output layer. The weights of the hidden layers and the output layer are initialized normally and using Glorot uniform initializer (Glorot & Bengio, 2010). MSE is used as the loss/cost function. The NADAM (Dozat, 2016) optimizer, which is the same as Adam (Kingma & Ba, 2015) optimizer with Nesterov momentum, is used for the stochastic optimization of the DNNs. Proper learning rates of the optimizer are selected and mini-batches of different sizes are used for each SQM. Dropouts (Srivastava et al., 2014) of different values are used after each hidden layer to minimize the risk of overfitting and to speed up the DNN training. The dataset is standardized by taking the mean and scaling to unit variance. Table 5 presents the choices of parameters and hyper-parameters of DNN-based SQMs trained and optimized for each noise class individually.

5.2 Training stage

For training the DNN-based SQMs, the data of each noise class is divided into 80:20 ratio, that is, 80% data is used for training and 20% for testing. The entire training data is used for optimizing the weights of each DNN model.

5.3 Testing stage

During testing, the test speech samples of each noise class are passed to the trained classifier (XGBoost) for correctly identifying the noise class of that test sample. After identifying its noise class, the test sample is passed to the corresponding trained and optimized SQM for obtaining the optimized objective speech quality predictions (MOS-LQO). In addition, all test speech samples together are also passed through the trained classifier for detecting its noise classes and then corresponding SQMs to obtain the objective speech quality predictions (MOS-LQO).

6 Results and discussion

This section presents the numerical results, which fit into our proposed metric and offers the solution to the gap discovered in the literature. It also discusses the results of each building block to show-case the effectiveness of our proposed metric.

6.1 Noise-class classifier response

The Precision, Recall and F-score of the XGBoost classifier for each noise class are presented in Table 6 and the G-mean (balanced accuracy) of the XGBoost classifier for each noise class is presented in Table 7. It can be seen from both Tables that the classifier has average test accuracy and balanced accuracy of 82%. Therefore, XGBoost classifier is used further to develop the complete proposed metric.

6.2 Voice activity detection response

The performance of the WS VAD (Jaiswal, 2022) is measured on noisy speech samples of each noise class of the NOIZEUS speech corpus (see Sect. 3) at two SNRs, that is, 5dB and 10 dB and compared to the ground truth (GT) VAD, energy (E) VAD and weighted energy[7] (WE) VAD. The Precision, Recall and F-score are calculated for each noise class and the F-score is presented in Table 8. It can be noticed that the E and WE VAD are inaccurately detecting the voiced segments as compared to the

[7] Energy and Weighted energy VAD are developed by extracting the energy features of the speech samples in our previous study (Jaiswal, 2022).

Table 5 Parameters and hyper-parameters of each DNN-based speech quality metric (SQM)

Parameters default to all SQMs				
Input layer neurons	10			
Number of hidden layers	3			
Output layer neurons	1			
Output layer activation function	Linear			
Weight initialization of 1st hidden layer	Normal			
Weight initialization of 2nd and 3rd hidden layers	Glorot uniform			
Weight initialization of output layer	Glorot uniform			
Parameters for individual SQMs				
	Airport SQM	Exhibition SQM	Restaurant SQM	Station SQM
Neurons in each hidden layer	13, 141, 10	13, 3, 10	13, 10, 10	13, 3, 10
Hidden layers activation function	ReLU, Linear, Linear	ReLU, ReLU, ReLU	ReLU, ReLU, ReLU	ReLU, ReLU, ReLU
Dropout after each hidden layer	0.3, 0.3, 0.3	0.2, 0.2, 0.2	0.4, 0.5, 0	0.2, 0.2, 0.2
Optimizer	NADAM	NADAM	NADAM	NADAM
Learning rate	0.0008	0.0008	0.0008	0.0008
Number of epochs	1900	1900	1000	1900
Size of mini-batch	10	10	4	10
Regularizer of RidgeCV	0.01	0.01	0.01	–
Max. no. of iterations of LassoCV	–	–	–	10,000

Table 6 Precision, Recall and F-score of XGBoost classifier for each noise class along with average score

	Airport	Exhibition	Restaurant	Station	Average
Precision	0.89	1	1	1	0.97
Recall	0.67	0.75	0.67	0.75	0.71
F-score	0.76	0.86	0.80	0.86	0.82

Table 7 G-mean of XGBoost classifier for each noise class along with average score

	Airport	Exhibition	Restaurant	Station	Average
XGBoost	0.78	0.86	0.81	0.86	0.82

GT VAD, resulting in poor accuracy. However, the WS VAD performs excellent in correctly detecting the speech components for each noise class, resulting in high accuracy. Moreover, while testing all 240 noisy speech samples together, the WS VAD shows an accuracy of 94.7%. Therefore, the WS VAD is highly robust to pre-process the speech samples and can be integrated further for developing the complete proposed metric.

6.3 Speech quality metric response

Table 9 presents the training and the testing errors while training a collection of DNNs to develop SQMs associated with each noise class. It can be noticed that the accuracy, that is, the test MSE of each SQM is comparable to its counter training MSE. All quality predictions of test samples are estimated with a small error in the range of 0.01 to 0.04, that is, 1 to 4%. It reflects that the MSE between the training and the testing quality estimates of the individual SQM is very small. Hence, the results imply a better quality predictions for our SQMs.

Figure 8 shows the accuracy of each trained speech quality metric in terms of MSE. It can be easily visualized that the individual metric learning is smoother and it converges towards the local minima, that is, as the number of epochs increases, the training loss (MSE) decreases and becomes stable. The testing curve follows it, resulting in optimized accuracy. Therefore, these noise-class specific (context-specific) optimized SQMs can be integrated further for developing the complete proposed metric.

Table 9 Model learning for each speech quality metric (SQM)

	Airport SQM	Exhibition SQM	Restaurant SQM	Station SQM
Train MSE	0.032	0.026	0.014	0.028
Test MSE	0.039	0.044	0.031	0.035

6.4 Overall proposed metric response

This section presents Pearson correlation "ρ_p", Spearman correlation "ρ_s" and RMSE between the objective quality scores (MOS-LQO) obtained for each noise class and their corresponding subjective listener quality scores (MOS-LQS). Three different scenarios are tested to investigate the effectiveness of our proposed metric.

6.4.1 Scenario A: proposed metric without noise-class classifier

In this scenario, the noise-class classifier is removed, that is, the noise class of the input noisy speech sample is not detected. The input noisy speech samples are just processed by the WS VAD and then the processed samples are injected to the P.563 metric to obtain the objective quality scores (MOS-LQO). Table 10 presents ρ_p, ρ_s and RMSE for each noise class in this scenario.

It can be observed from Table 10 that the ρ_p and ρ_s of each noise class including all noise classes grouped together are very poor, that is, both correlation values range in between 26 and 52% only. When all noise classes are grouped together, then the correlation values of ρ_p and ρ_s are only 0.428 and 0.432, respectively, which are very less. Moreover, the RMSE is also very high, ranging between 83 and 97%. This implies that Scenario A is not suitable for measuring and monitoring quality of speech in real-time. Moreover, the results also direct that classifying the noise class may result in a better speech quality prediction which is incorporated in our proposed metric (Scenario C).

Table 8 F-score of VAD for each noise class along with all speech samples from each noise class together

VAD	Airport		Exhibition		Restaurant		Station		All
	5 dB	10 dB	5 dB	10 dB	5 dB	10 dB	5 dB	10 dB	Samples
GT	1	1	1	1	1	1	1	1	1
E	0.300	0.268	0.301	0.264	0.304	0.269	0.295	0.263	0.270
WE	0.587	0.520	0.562	0.513	0.597	0.510	0.565	0.508	0.553
WS	0.909	0.915	0.965	0.944	0.921	0.928	0.911	0.891	0.947

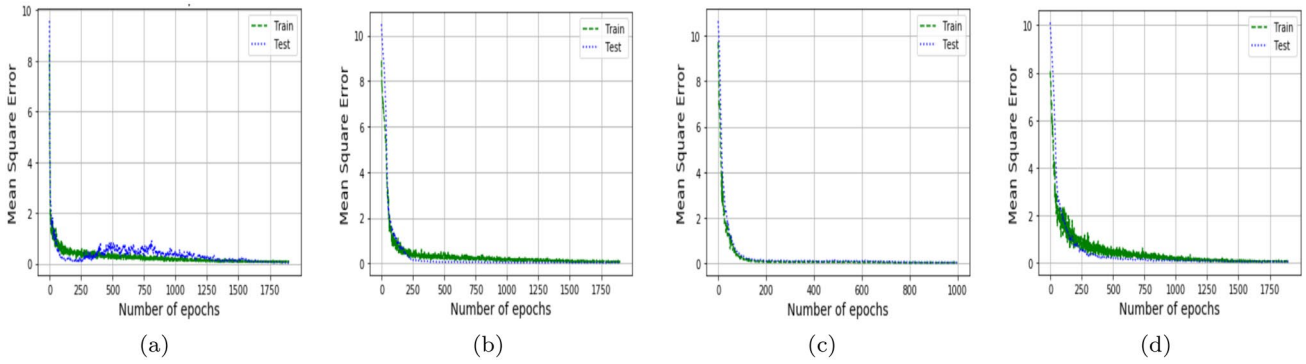

Fig. 8 Accuracy in terms of MSE for each speech quality metric **a** Airport SQM, **b** Exhibition SQM, **c** Restaurant SQM, and **d** Station SQM

Table 10 ρ_p, ρ_s and RMSE for each noise class with a grouped results for all noise classes without Noise-class Classifier (Scenario A)

Noise class →	Airport	Exhibition	Restaurant	Station	All
ρ_p	0.407	0.524	0.268	0.472	0.428
ρ_s	0.391	0.507	0.286	0.488	0.432
RMSE	0.972	0.837	0.923	0.943	0.920

6.4.2 Scenario B: proposed metric without both noise-class classifier and VAD

In this scenario, both noise-class classifier and VAD are removed, that is, neither the noise class of the input noisy speech samples are detected nor the input noisy speech samples are processed by the VAD. In other words, the input noisy speech samples are directly injected to the P.563 metric to obtain the objective quality scores (MOS-LQO). Table 11 presents ρ_p, ρ_s and RMSE for each noise class in this scenario.

It can be observed from Table 11 that the ρ_p and ρ_s of each noise class including all noise classes grouped together are very poor, that is, both correlation values range in between 35 and 52% only. When all noise classes are grouped together, then the correlation values of ρ_p and ρ_s are only 0.440 and 0.448, respectively, which are very less. Moreover, the RMSE is also very high, ranging between 80 and 94%. However, the results obtained in this scenario (Scenario B) is slightly better than the results obtained in Scenario A. This implies that Scenario B is also not suitable for measuring and monitoring quality of speech in real-time. Moreover, the results also direct that classifying the noise class and pre-processing the speech samples via a VAD may result in a better speech quality prediction. Therefore, both noise-class classifier and VAD are incorporated in our proposed metric (Scenario C).

Table 11 ρ_p, ρ_s and RMSE for each noise class with a grouped results for all noise classes without both Noise-class Classifier and VAD (that is, only P.563 is present) (Scenario B)

Noise class →	Airport	Exhibition	Restaurant	Station	All
ρ_p	0.399	0.524	0.350	0.483	0.440
ρ_s	0.426	0.494	0.366	0.492	0.448
RMSE	0.947	0.809	0.861	0.942	0.892

Table 12 ρ_p, ρ_s and RMSE for each noise class with a grouped results for all noise classes in our main proposed metric (Scenario C)

Noise class →	Airport	Exhibition	Restaurant	Station	All
ρ_p	0.948	0.817	0.768	0.973	0.931
ρ_s	0.637	0.750	0.857	0.872	0.928
RMSE	0.293	0.334	0.234	0.325	0.232

6.4.3 Scenario C: proposed metric

This is our main proposed noise class-awared speech quality prediction metric scenario where both noise-class classifier and VAD are present, that is, the noise class of the input noisy speech samples are firstly detected, and then processed with the WS VAD to segregate silences, thereafter the processed speech samples are injected to the P.563 metric to obtain objective speech quality scores (MOS-LQO). Table 12 presents ρ_p, ρ_s and RMSE for each noise class in this scenario.

It can be observed from the results of our proposed metric (Scenario C), presented in Table 12, that the ρ_p and ρ_s of station class is highest, followed by the airport class. The exhibition and restaurant class also have better ρ_p but less than the airport class. The restaurant class has better ρ_s, followed by the exhibition and the airport class. The

correlation value in terms of ρ_p ranges in between 76 and 97%. Similarly, the correlation value in terms of ρ_s ranges in between 63 and 92%. When all noise classes are grouped together, then the correlation values of ρ_p and ρ_s are 0.931 and 0.928, respectively, which are better. In addition, the RMSE of the restaurant class is smallest among other noise classes. The RMSE varies in a range of 23–33%, which is in the acceptable range. Overall, our proposed metric performs outstanding, giving around 93% accuracy in terms of both correlations and lowest RMSE when tested with all test samples taken from each noise class together. This implies that incorporation of both noise-classifier and a VAD in our proposed metric (Scenario C) is justified. As a result, our proposed metric is suitable for measuring and monitoring speech quality in real-time.

6.4.4 Comparison of all scenarios

It can be observed from Tables 10, 11 and 12 that the ρ_p, ρ_s and RMSE of our proposed SQM (Scenario C) are better than both Scenario A and Scenario B for each noise class and for all noise classes tested together. The ρ_p value (0.931) is highest (see Table 12) in our proposed metric (Scenario C) as compared to both Scenario A (0.428, see Table 10) and Scenario B (0.440, see Table 11) when tested with all noise classes together. Similarly, the ρ_s value (0.928) is highest (see Table 12) in our proposed metric (Scenario C) as compared to both Scenario A (0.432, see Table 10) and Scenario B (0.448, see Table 11) when tested with all noise classes together. In addition, the RMSE value (0.232) is lowest (see Table 12) in our proposed metric (Scenario C) as compared to both Scenario A (0.920, see Table 10) and Scenario B (0.892, see Table 11) when tested with all noise classes together. Moreover, both correlation values in our proposed metric (Scenario C) are around 53% higher as compared to both Scenario A and Scenario B. Similarly, our proposed metric (Scenario C) has around 73% lower RMSE as compared to both Scenario A and Scenario B. This reflects that our proposed metric is performing outstanding as compared to these two different baseline scenarios. The deployment of both pipelines, that is, the noise-class classifier for detecting the noise class of the input noisy speech sample and the WS VAD for identifying and separating the voiced segments, are the key components of our proposed speech quality metric. In particular, it is the sensitivity of noise class that is leading to the better results in our proposed metric.

6.4.5 Plots of correlations and RMSE

In order to clearly visualize and compare the obtained results, Fig. 9 presents the correlations and RMSE of our proposed metric (Scenario C) with two different baselines (Scenario A and Scenario B). It can be seen that our proposed metric is having higher correlations and lower RMSE in case of each noise class and all noise classes together, showing superior performance.

6.4.6 Scatter plot between subjective and objective MOS

To explore the impact of our proposed metric, the scatter plot between the objective quality predictions obtained from each speech quality metric which is trained for each noise class individually (denoted by MOS-LQO) and the corresponding subjective quality predictions (denoted by MOS-LQS) is depicted in Fig. 10. A good correlation can be observed with the test samples of each noise class here.

6.4.7 Comparison of execution time

One of the important figures of merit of the speech quality assessment metric for real-time speech quality measuring and monitoring is the processing time. In this regard, Table 13 presents the difference in computational complexity between our proposed metric and the P.563 metric while executing a single test speech sample. It can be seen that our proposed metric is faster, saving computational time.

The overall simulation results of the proposed SQM show that the deep learning-based approach has a great advantage when the speech signals are distorted by various types of noise degradations in the surroundings. It demonstrates that DNNs have the great potential to remember and analyze the complicated characteristics of speech signals. Therefore, we believe that our proposed SQM can be deployed by the internet service providers for measuring and monitoring real-time speech quality in the environments where the speech quality is degraded due to the presence of different types of background noises and then QoE-aware management actions can be taken in order to react and maintain the end-user QoE levels.

7 Conclusions and future work

This paper proposes a context-aware speech quality assessment metric for measuring and monitoring the real-time speech quality of voice communication systems under noisy environments. The first component of the proposed metric is a noise-class classifier which uses speech enhancement algorithms in conjunction with P.563 metric to compute speech quality scores (MOS scores) to be used as input feature vector for training the classifier to detect the noise class (context) of the input noisy speech signal. The second component is a VAD, which pre-processes the speech samples

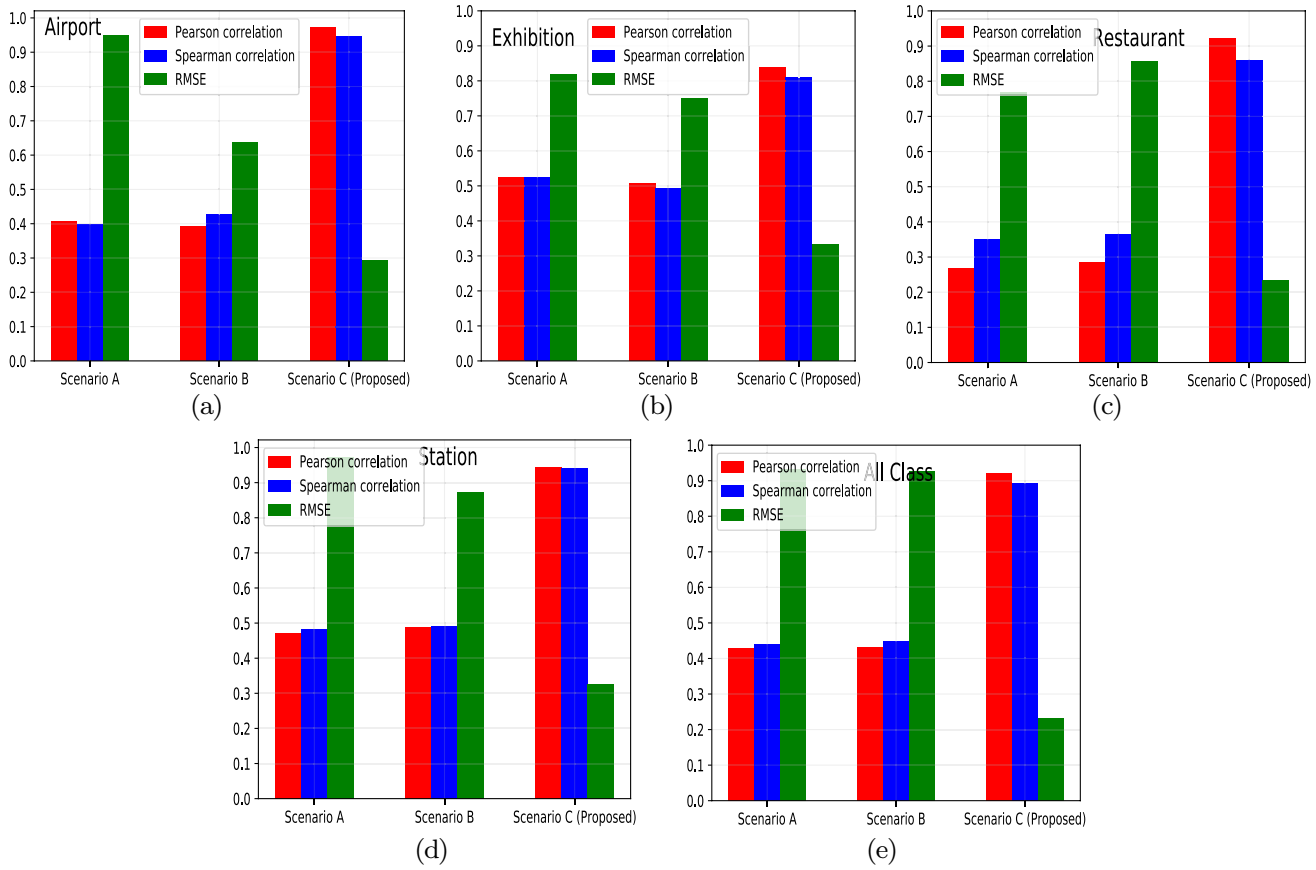

Fig. 9 Correlations and RMSE for two different baselines (Scenario A and Scenario B) and our proposed metric (Scenario C): **a** Airport class, **b** Exhibition class, **c** Restaurant class, **d** Station class, and **e** All noise classes together

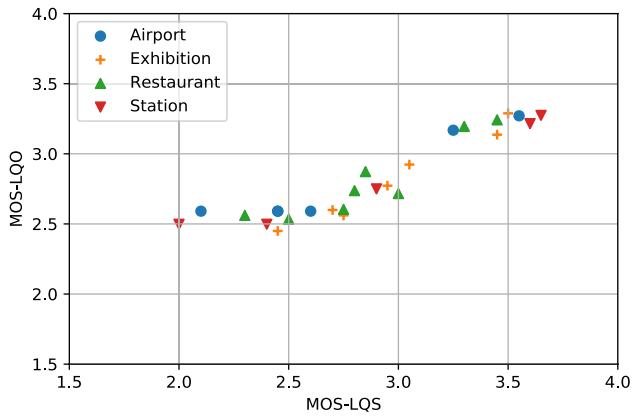

Fig. 10 Subjective vs. objective quality predictions of each noise class

Table 13 Execution time for a single speech sample

	P.563	Proposed metric	Difference
Time (s)	0.682	0.154	0.528

to identify and segregate the voiced segments. The third component is noise class-specific (context-specific) SQMs, which are developed by training and optimizing DNNs for each noise class individually. The noise class of the input test speech sample is identified by the noise-class classifier and then the corresponding SQM is activated by the classifier to obtain the optimized speech quality predictions (MOS scores). Results illustrate that the correlation between the subjective and the objective quality predictions is high and the RMSE is less for each noise class individually and when all noise classes are tested together, for our proposed metric as compared to the metric where the noise class is not detected before the prediction of speech quality. The proposed metric also takes less amount of time to execute the speech sample. This indicates that the proposed metric estimates the perceived quality of speech sequences with better accuracy. The proposed speech quality assessment metric is computationally efficient and presents a significant advantage over the SQMs present in the literature where the speech quality is predicted directly without identifying the noise class explicitly.

The proposed metric is not limited to narrow-band speech signals. The narrow-band nature of the test data is used for the developed model. However, the same strategy can be applied with wide-band speech signals in future. We also estimate that extending the dataset further with varieties of noise classes will produce even more better results. Developing a VoIP conversational dataset and its subjective quality rating to validate the proposed metric, since there is no publicly available VoIP conversational dataset in the literature, are the part of future considerations.

Funding Open access funding provided by University of Agder.

Declarations

Conflict of interest The authors would like to declare that there is no conflict of interest with this manuscript.

References

Avila, A. R., Gamper, H., Reddy, C., Cutler, R., Tashev, I., & Gehrke, J. (2019). Non-intrusive speech quality assessment using neural networks. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2019 (pp. 631–635).

Belarouci, S., & Chikh, M. A. (2017). Medical imbalanced data classification. *Advances in Science, Technology and Engineering Systems Journal, 2*(3), 116–124.

Bergstra, J. A., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(2), 281–305.

Bergstra, J. A., & Middelburg, C. (2003). *ITU-T Recommendation G.107: The E-Model, a computational model for use in transmission planning*. International Telecommunication Union.

Bruhn, S., Grancharov, V., & Kleijn, W. B. (2012). Low-complexity, non-intrusive speech quality assessment. US Patent, 8,195,449.

Catellier, A. A., & Voran, S. D. (2020). Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 331–335).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016 (pp. 785–794).

Chinen, M., Lim, F. S., Skoglund, J., Gureev, N., O'Gorman, F., & Hines, A. (2020). ViSQOL v3: An open source production ready objective speech and audio metric. In *Twelfth international conference on quality of multimedia experience*, 2020 (pp. 1–6). IEEE.

Chowdhury, A., Yang, J., & Drineas, P. (2018). An iterative, sketching-based framework for ridge regression. In *International conference on machine learning*, 2018 (pp. 989–998).

Cohen, I. (2002). Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *IEEE Signal Processing Letters, 9*(4), 113–116.

Das, N., Chakraborty, S., Chaki, J., Padhy, N., & Dey, N. (2020). Fundamentals, present and future perspectives of speech enhancement. *International Journal of Speech Technology, 24,* 1–19.

Deligiannidis, L., & Arabnia, H. R. (2014). *Emerging trends in image processing, computer vision and pattern recognition*. Morgan Kaufmann.

Dimitrakopoulos, G. N., Vrahatis, A. G., Plagianakos, V., & Sgarbas, K. (2018). Pathway analysis using XGBoost classification in biomedical data. In *Proceedings of the 10th Hellenic conference on artificial intelligence*, 2018 (pp. 1–6).

Dozat, T. (2016). Incorporating Nesterov momentum into Adam. In *4th International conference on learning representations (ICLR)*, 2016.

Drummond, C., & Holte, R. C. (2003). C 4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *20th International conference on machine learning (ICML) workshop on learning from imbalanced data sets*, 2003.

Dubey, R. K., & Kumar, A. (2013). Non-intrusive speech quality assessment using several combinations of auditory features. *International Journal of Speech Technology, 16*(1), 89–101.

Dubey, R. K., & Kumar, A. (2015). Comparison of subjective and objective speech quality assessment for different degradation/noise conditions. In *IEEE international conference on signal processing and communication*, 2015 (pp. 261–266).

Dubey, R. K., & Kumar, A. (2017). Non-intrusive speech quality estimation as combination of estimates using multiple time-scale auditory features. *Digital Signal Processing, 70,* 114–124.

Eisen, M., Zhang, C., Chamon, L. F., Lee, D. D., & Ribeiro, A. (2018). Online deep learning in wireless communication systems. In *52nd Asilomar conference on signals, systems, and computers (ACSSC)*, 2018 (pp. 1289–1293). IEEE.

Engelberg, S. (2008). *Digital signal processing: An experimental approach*. Springer.

Ephraim, Y. (1992). Statistical-model-based speech enhancement systems. *Proceedings of the IEEE, 80*(10), 1526–1555.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(6), 1109–1121.

Falk, T. H., & Chan, W. Y. (2006). Single-ended speech quality measurement using machine learning methods. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 1935–1947.

Falk, T. H., Zheng, C., & Chan, W. Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(7), 1766–1774.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced datasets* (Vol. 11). Springer.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(4), 1189–1232.

Fu, S. W., Tsao, Y., Hwang, H. T., & Wang, H. M. (2018). Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM. In *Interspeech*, 2018.

Fukuda, T., Ichikawa, O., & Nishimura, M. (2018). Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition. *Speech Communication, 98,* 95–103.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th*

international conference on artificial intelligence and statistics, 2010 (pp. 249–256).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative nets. In *Advances in neural information processing systems*, 2014 (pp 2672–2680).

Grancharov, V., Zhao, D. Y., Lindblom, J., & Kleijn, W. B. (2006). Low-complexity, non-intrusive speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 1948–1956.

Gustafsson, H., Nordholm, S. E., & Claesson, I. (2001). Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Transactions on Speech and Audio Processing, 9*(8), 799–807.

Hines, A., Gillen, E., & Harte, N. (2015a). Measuring and monitoring speech quality for voice over IP with POLQA, ViSQOL and P.563. In *INTERSPEECH*, 2015, Dresden, Germany.

Hines, A., Skoglund, J., Kokaram, A. C., & Harte, N. (2015). ViSQOL: An objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing, 1,* 1–18.

Hirsch, H. G., & Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000—Automatic Speech Recognition: Challenges for the new millennium, ISCA tutorial and research workshop (ITRW)*, 2000, Paris, France.

Holub, J., Avetisyan, H., & Isabelle, S. (2017). Subjective speech quality measurement repeatability: Comparison of laboratory test results. *International Journal of Speech Technology, 20*(1), 69–74.

Hu, Y., & Loizou, P. C. (2003). A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Transactions on Speech and Audio Processing, 11,* 334–341.

Hu, Y., & Loizou, P. C. (2004). Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE Transactions on Speech and Audio Processing, 12*(1), 59–67.

Hu, Y., & Loizou, P. C. (2006). Subjective comparison of speech enhancement algorithms. In *IEEE international conference on acoustics speech and signal processing*, Vol. 1, (pp. 153–156).

Hu, Y., & Loizou, P. C. (2007). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(1), 229–238.

ITU. (1996). *ITU-T Recommendation P.800: Methods for subjective determination of transmission quality*. ITU.

ITU. (1998). *ITU-T coded-speech database: Series P*, Supplement 23. ITU.

ITU. (2004). *ITU-T recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications*. ITU.

ITU. (2011). *ITU-T recommendation P.863: Perceptual objective listening quality assessment (POLQA)*. ITU.

Jahromi, H. Z., Hines, A., & Delanev, D. T. (2018). Towards application-aware networking: ML-based end-to-end application KPI/QoE metrics characterization in SDN. In *Tenth international conference on ubiquitous and future networks (ICUFN)*, 2018 (pp. 126–131).

Jain, R., Damoulas, T., & Kontokosta, C. (2014). Towards data-driven energy consumption forecasting of multi-family residential buildings: feature selection via the lasso. In *Computing in civil and building engineering*, 2014 (pp. 1675–1682).

Jaiswal, R. (2022). Performance analysis of voice activity detector in presence of non-stationary noise. In *Proceedings of the 11th international conference on robotics, vision, signal processing and power applications (RoViSP)*, 2022 (pp. 59–65). Springer.

Jaiswal, R., & Hines, A. (2018). The sound of silence: How traditional and deep learning based voice activity detection influences speech quality monitoring. In *26th Irish conference on artificial intelligence and cognitive science (AICS)*, 2018 (pp. 174–185).

Jaiswal, R., & Hines, A. (2020). Towards a non-intrusive context-aware speech quality model. In *31st Irish signals and systems conference*, 2020 (pp. 1–5). IEEE.

Kamath, S., & Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *IEEE international conference on acoustics speech and signal processing*, Vol. 4, (pp. 4160–4164).

Kim, D. S., & Tarraf, A. (2007). ANIQUE+: A new American national standard for non-intrusive estimation of narrow-band speech quality. *Bell Labs Technical Journal, 12*(1), 221–236.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International conference on learning representations*, 2015.

Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems, 53*(3), 551–577.

Malfait, L., Berger, J., & Kastner, M. (2006). P.563—The ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 1924–1934.

Meijer, R. J., & Goeman, J. J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal, 55*(2), 141–155.

Mittal, U., & Phamdo, N. (2000). Signal/noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Transactions on Speech and Audio Processing, 8*(2), 159–167.

Möller, S., Chan, W. Y., Côté, N., Falk, T. H., Raake, A., & Wältermann, M. (2011). Speech quality estimation: Models and trends. *IEEE Signal Processing Magazine, 28*(6), 18–28.

Moore, A. H., Parada, P. P., & Naylor, P. A. (2017). Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures. *Computer Speech and Language, 46,* 574–584.

Ooster, J., Huber, R., & Meyer, B. T. (2018). Prediction of perceived speech quality using deep machine listening. In *INTERSPEECH*, 2018 (pp. 976–980).

Ramirez, J., Górriz, J. M., & Segura, J. C. (2007). Voice activity detection: Fundamentals and speech recognition system robustness. *Robust Speech Recognition and Understanding, 6*(9), 1–22.

Reddy, M. K., Helkkula, P., Keerthana, Y. M., Kaitue, K., Minkkinen, M., Tolppanen, H., et al. (2021). The automatic detection of heart failure using speech signals. *Computer Speech and Language, 69,* 101205.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. In *IEEE international conference on acoustics, speech, and signal processing*, 2001, Vol. 2, (pp. 749–752).

Saleem, N., & Khattak, M. I. (2019). A review of supervised learning algorithms for single channel speech enhancement. *International Journal of Speech Technology, 22*(4), 1051–1075.

Scalart, P., et al. (1996). Speech enhancement based on a priori signal to noise estimation. In *IEEE international conference on acoustics, speech, and signal processing*, 1996, Vol. 2, (pp. 629–632).

Shami, M., & Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication, 49*(3), 201–212.

Sharma, D., Wang, Y., Naylor, P. A., & Brookes, M. (2016). A data-driven non-intrusive measure of speech quality and intelligibility. *Speech Communication, 80,* 84–94.

Shome, N., Laskar, R. H., & Das, D. (2019). Reference free speech quality estimation for diverse data condition. *International Journal of Speech Technology, 22*(3), 585–599.

Singh, J., & Singh, J. (2021). A survey on machine learning-based malware detection in executable files. *Journal of Systems Architecture, 112,* 101861.

Soni, M. H., & Patil, H. A. (2021). Non-intrusive quality assessment of noise-suppressed speech using unsupervised deep features. *Speech Communication, 130,* 27–44.

Srivastava, N., et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Sun, H., Chen, X., Shi, Q., Hong, M., Fu, X., & Sidiropoulos, N. D. (2017). Learning to optimize: Training deep neural networks for wireless resource management. In *18th IEEE international workshop on signal processing advances in wireless communications*, 2017 (pp. 1–6).

Wang, J., Shan, Y., Xie, X., & Kuang, J. (2019). Output-based speech quality assessment using autoencoder and support vector regression. *Speech Communication, 110,* 13–20.

Yang, H., et al. (2016). Parametric-based non-intrusive speech quality assessment by deep neural network. In *IEEE international conference on digital signal processing (DSP)*, 2016 (pp. 99–103).

Ye, H., Li, G. Y., & Juang, B. H. (2017). Power of deep learning for channel estimation and signal detection in OFDM systems. *IEEE Wireless Communications Letters, 7*(1), 114–117.

Ye, H., Li, G. Y., Juang, B. H. F., & Sivanesan, K. (2018). Channel agnostic end-to-end learning based communication systems with conditional GAN. In *IEEE GLOBECOM Workshop*, 2018 (pp. 1–5).