IET Communications

Research Article

IET Journals
The Institution of Engineering and Technology

NIML: non-intrusive machine learning-based speech quality prediction on VoIP networks

ISSN 1751-8628
Received on 23rd May 2018
Revised 28th February 2019
Accepted on 4th July 2019
E-First on 30th August 2019
doi: 10.1049/iet-com.2018.5430
www.ietdl.org

Rami S. Alkhawaldeh[1] ✉, Saed Khawaldeh[2,3,4,5], Usama Pervaiz[2,3,4,5], Moatsum Alawida[6], Hamzah Alkhawaldeh[7]

[1]*Department of Computer Information Systems, The University of Jordan, Aqaba 77110, Jordan*
[2]*Erasmus+ Joint Master Program in Medical Imaging and Applications, University of Burgundy, France*
[3]*Erasmus+ Joint Master Program in Medical Imaging and Applications, University of Cassino, Italy*
[4]*Erasmus+ Joint Master Program in Medical Imaging and Applications, University of Girona, Spain*
[5]*Sensor Informatics and Medical Technology Group, Department of Electrical Engineering and Automation, Aalto University, Finland*
[6]*School of Computer Sciences, University Sains Malaysia, 11800 USM, Pulau Pinang, Malaysia*
[7]*King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman 11941, Jordan*
✉ *E-mail: r.alkhawaldeh@ju.edu.jo*

Abstract: Voice over Internet Protocol (VoIP) networks have recently emerged as a promising telecommunication medium for transmitting voice signal. One of the essential aspects that interests researchers is how to estimate the quality of transmitted voice over VoIP for several purposes such as design and technical issues. Two methodologies are used to evaluate the voice, which are subjective and objective methods. In this study, the authors propose a non-intrusive machine learning-based (NIML) objective method to estimate the quality of voice. In particular, they build a training set of parameters – from the network and the voice itself – along with the quality of voices as labels. The voice quality is estimated using the perceptual evaluation of speech quality (PESQ) method as an intrusive algorithm. Then, the authors use a set of classifiers to build models for estimating the quality of the transmitted voice from the training set. The experimental results show that the classifier models have a valuable performance where Random Forest model has superior results compared to other models of precision 94.1%, recall 94.2%, and receiver operating characteristic area 99.2% as evaluation metrics.

1 Introduction

In the last decades, transmitting voice over data-centric VoIP networks has an intrinsic attention in the telecommunication industry [1]. Instead of transmitting the voice over a dedicated path in public switched telephone networks (PSTN), the voice data are sent over a VoIP network using a set of processes [2]. The processes include encoding, packetising, and then sending the packets over internet protocol (IP) using non-guaranteed user datagram protocol (UDP). On the other side, the process is reversed to aggregate the packets into received voice as shown in Fig. 1.

The reasons for using the data-centric networks include [3]: (i) saving money in comparison with PSTN networks, (ii) having one information service department of a group of administrative staff, and (iii) constructing a standard unified network for voice and data traffic that should be improved for advanced services. Although VoIP is used for propagating real-time traffic, the received voice is degraded due to several impairments occurred for packets during their trip in the network. For instance, the transmitted voice packets might be lost in the network due to congestion in the IP network, delay in decoding processes, or delay variance (or jitter) of receiving the packets. In jitter case, the delay occurs over time from point-to-point when network congestion, improper queuing occurs, or low bandwidth situations. Hence, the impairments decrease the quality of voice in IP networks when it is compared to the quality of voice in PSTN networks. This motivates to discover techniques that assess the quality of VoIP traffic for legal, commercial, and technical reasons [4]. Users in VoIP networks interests with a high-quality voice as quality service [5]. Since in order to provide such quality of service, VoIP service providers consider measuring voice quality as one of the essential processes for achieving user satisfaction by improving the quality of the network.

There are two methodologies to evaluate the quality of voice in VoIP networks that depend on the involvement of users in assessment, which are subjective and objective methods [6]. In subjective methods, the users estimate the voice quality using calibrated laboratory under standard conditions, while in the objective methods computational algorithms are used to evaluate the quality of voice without user interference. Although subjective methods result in a highly effective approach to evaluate the quality, it is difficult to be achieved due to cost, time consuming, and trustworthiness of users in giving their opinion. The objective methods are used instead that depend on several parameters to evaluate the quality of voice. There are two categories of objective methods depending on availability of the original voice signal,

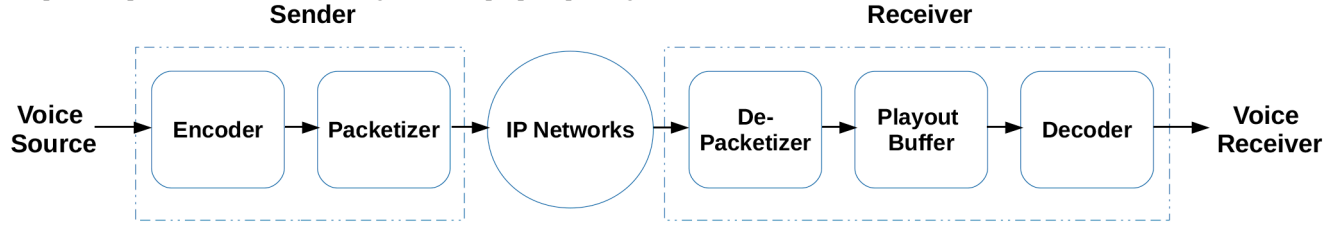

Fig. 1 *VoIP processes [2]*

namely, intrusive and non-intrusive methods. Intrusive methods use original and degraded voice signals to measure distortion as a reference for quality evaluation [6]. Non-intrusive methods use only the degraded signal to estimate the quality whether using some parameters from the network [7] or features from signal processing techniques [8]. Despite the accuracy of intrusive methods, sometimes it is hard to obtain the original voice for evaluation. Hence, non-intrusive methods are more promising and realistic techniques in VoIP networks.

In this paper, we propose a non-intrusive machine learning-based objective method (NIML) using parameters from the network and from the transmitted signal voice as features. The proposed approach combines the characteristics of intrusive and non-intrusive methods in extracting features in order to alleviate their problems. In particular, the NIML method is a machine learning approach that is built on a training set of features and labels. The features include encoding/decoding type, the voice (speech) gender, language, packet loss, and burst ratio, whereas the labels include mean opinion score (MOS) that are converted from perceptual evaluation of speech quality (PESQ) scores. Our contributions are summarised as follows:

• We investigate the effect of parameters for voice quality assessment used in the training set and their applicability in building the classifiers.

RQ1: How effect is using the selected parameters on the quality of voice?

• We study a set of classifiers that are selected from different families of machine learning, and estimate their robustness in evaluating the voice quality in comparison with the PESQ intrusive technique.

RQ2: What is the robustness of using different learners in evaluating the voice quality in the VoIP networks?

This paper is organised as follows. Section 2 discusses the quality of service and the methods that used to estimate the voice quality. Section 3 presents the NIML objective method voice quality assessments along with the parameters that are used for building a training set. Section 4 argues the parameters settings and the simulation for extracting their values, in addition to the machine learning and evaluation techniques that are used for validating the NIML approach. Finally, Section 5 summarises our approach and the results along with discussing the research questions, in addition to future works.

2 Related work

Several methods have been emerged for evaluating the quality of transmitted voice over VoIP networks. This section clarifies a set of evaluating approaches that are classified as subjective and objective approaches.

2.1 Subjective methods

Subjective assessment methods need a panel of users to perform the subjective tests of voice quality (test subjects) by listening to a set of recorded or live conversational speech samples in different network conditions under special laboratory settings. Each subject gives his/her opinion on voice quality sample by assigning a score value to each sample ranging from 1 refer to bad quality up to 5 refer to excellent quality [9]. These scores result in a MOS metric that is estimated by averaging the scores obtained from the panel of users for a given sample such as the dominant method absolute category rating (ACR). The subjective assessment methods are more accurate, useful, and reliable in measuring the perceived voice quality than the objective methods. The reason is that it reflects the opinion of human assessors at receiver end on voice quality. Therefore, the subjective methods are crucial as a benchmark for evaluating the objective methods. Nevertheless, it has several disadvantages such as [10]: (i) it needs special laboratory conditions for testing, (ii) it is expensive to perform the test because it needs many materials and resources to perform this operation, (iii) cost time to perform the test (time consuming), and (iv) finally it is not suitable in real time or online applications as in non-managed networks such as internet. There are two techniques of subjective assessment methods: conversional test (two-way test) and listening test (one-way test) [11].

2.2 Objective methods

As subjective speech quality tests are hard-to-conduct, consequently objective tests received great interest from researchers and engineers for a long time, who have attempted to evolve and develop new algorithmic, mathematical, and computational machine models. These models automatically evaluate the transmitted speech quality over the IP-based networks that replace the human panel. The aim is to predict MOS values that are as close as possible to the rating obtained from a subjective test and to make up or circumvent the limitations of subjective assessment methods. Objective measures can be classified into two classes: intrusive and non-intrusive; where intrusive measures, often referred to as input-to-output measures, their measurements depend on comparing the original (clean or input) speech signal with the degraded (distorted or output) speech signal which are also comparison-based methods such as signal-to-noise ratio [12], bark spectral distortion [13], modified bark spectral distortion [14], enhanced modified bark spectral distortion [14], measuring normalising blocks [15], perceptual speech quality measure (PSQM) and PSQM+ [16], perceptual analysis/measurement system (PAMS) [17] and the dominant standard PESQ [18, 19]. On the other hand, non-intrusive measures (also known as output-based or single-ended or passive measures) use only the degraded signal without access to the original signal, non-intrusive methods are also divided into two subcategories, signal-based and parametric-based methods. Signal-based methods, as the name suggests, depend on digital signal processing that estimates the speech quality when the envelope of the speech signal suffered from degradation over time due to low-bit rate coding or transmission over noisy wireless links such as ITU-T recommendation P.563. On the other hand, parameter-based methods process the audio stream that is decoded after buffer playout to extract relevant information for estimating the voice quality such as the most prominent approaches E-model (ITU-T recommendation G.107) [20] and POLQA ITU-T recommendation P.863 (Perceptual Objective Listening Quality Assessment) [21].

3 Non-intrusive machine learning-based objective approach

Objective methods mainly employ the parametric and/or signal features of degraded voices for estimating the voice quality. The proposed NIML approach also creates a training set of features along with their labels (i.e voice quality) from which a classifier is built for assessing the voice quality of unknown future voices.

3.1 Problem formulation of NIML approach

In order to mathematically clarify the problem of estimating voice quality over VoIP networks, the formulating form is described as follows: assume that we have a set of original voices O_v and their degraded voices D_v that are generated by conducting the simulation on a set of network parameters such as packet loss (P_{loss}), Codec (C), and burst ratio (B_{ratio}). The voice quality is estimated by intrusive methods using O_v and D_v or by non-intrusive methods using only D_v as inputs. Our approach considers the problem as a set of features and their labels. The features are extracted from the network and transmitted voice, including language (L), gender (G) which are estimated from the voice itself using a set of techniques [22, 23] and C, B_{ratio}, and P_{loss} as a function extracted from the network $\phi(L, G, B_{\text{ratio}}, P_{\text{loss}})$. The labels present the PESQ_{MOS} values resulted from the PESQ method; corresponding to $o \in O_v$ and $d \in D_v$ voices as $\text{PESQ}(o, d) = \text{MOS}$. A classifier is built by

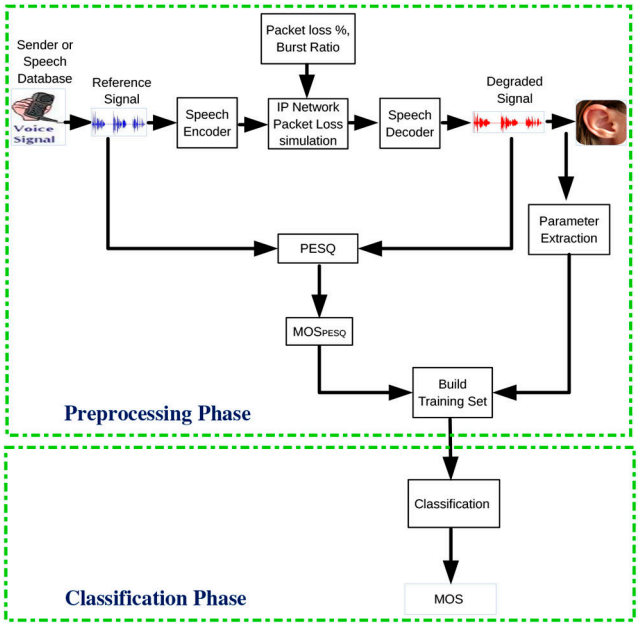

Fig. 2 *NIML voice quality prediction method*

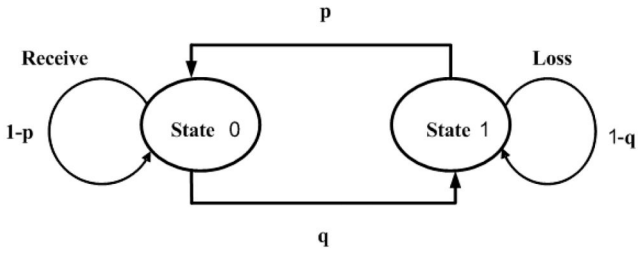

Fig. 3 *2-State Markov model*

mapping the set of features into their corresponding MOS value using the training set.

$$F(\phi(L, G, B_{\text{ratio}}, P_{\text{loss}})) \longrightarrow \text{PESQ}(o, d) \tag{1}$$

Fig. 2 shows our proposed approach and VoIP simulation. An original voice signal o is sent to the IP network after encoding. The simulation then distorts the voice with a percentage of P_{loss} and B_{ratio} values. Finally, the degraded voice signal d is decoded and then aggregated as packets voice at the destination end. The proposed approach builds a training set of features along with MOS_{PESQ} labels and constructs a classifier to predict the MOS_{PESQ} values.

3.2 Training set construction

The problem of voice-quality estimation depends essentially on the efficiency of constructed classifiers. We build a training set from a set of features derived from the network and the voice itself related with the MOS_{PESQ} values as labels estimated from the PESQ method. In more detail, the extracted features are as follows:

Packet loss (P_{loss}): The voice is encapsulated into voice packets that are transmitted over IP network using UDP. The UDP protocol does not guarantee the process of re-sending the lost packets in the network compared with transmission control protocol that does. This generated problems lead to the loss of some packets. These problems occurred due to some reasons such as (i) the hardware, like routers, sometimes cannot be able to meet bandwidth demands for receiving packets; (ii) many packets are lost at the receiver end due to the excessive delay or the interference observed on wireless links [24]. The packet loss in the packet-based network is inevitable, therefore packet loss still has an effect on the received quality.

To study the effect of packet loss on voice quality, a mathematical simulator model is needed. The state of packets at the receiver-end appears in found or lost state, which depends on the time of play-out. Since the Markov model is used to simulate the n-transition state of any system, we use it to simulate the 2-state transition of packet loss on packet-based networks [25]. The 2-state Markov model is shown in Fig. 3.

The 2-state Markov model uses two parameters p and q. The p parameter is the probability of transition from found state to loss state, while the q is the probability of transition from loss state to found state. We use the two parameters in an equation from ITU-T Recommendation G.107 [26] to simulate the packet loss on VoIP networks as shown in equation 2

$$\frac{1}{p+q} = \frac{\text{Packet\_Loss}}{p} = \frac{1-\text{Packet\_Loss}}{q} \tag{2}$$

The p and q parameters, in this equation, can be computed using packet loss. As such, the two parameters are used to simulate packet loss in a voice signal in VoIP networks.

Language (L): VoIP networks are essentially not restricted to a specific language where the architecture is independent of the transmitted voice signals. Hence, the network can use different languages of various effects on voice quality. To study such effects, standard artificial voices from ITU-T Recommendation P.50 [27] are created, which consists of 20 languages. The languages are American English (**A**), Arabic (**Ar**), British English (**B**), Chinese (**Ch**), Danish (**Da**), Dutch (**Du**), Finnish (**Fi**), French (**Fr**), German (**Ger**), Greek (**Gr**), Hindi (**Hin**), Hungarian (**Hu**), Italian (**Ita**), Japanese (**Ja**), Norwegian (**Nor**), Polish (**Po**), Portuguese (**Port**), Russian (**Ru**), Spanish (**Sp**), and Swedish (**Swe**). Each language has 32 voices of 16 for each gender. The artificial voice is a signal that is mathematically defined to reproduce the time and spectral characteristics of a human speech – over the bandwidth 100 Hz–8 kHz – which significantly affect the performances of linear and non-linear telecommunication systems. The following time and spectral characteristics of real speech are reproduced by the artificial voice [27]:

- long-term average spectrum.
- short-term spectrum.
- instantaneous amplitude distribution.
- voiced and unvoiced structure of speech waveform.
- syllabic envelope.

The artificial voice described in the Recommendation is mainly used for objective evaluation of speech processing systems and devices, in which a single-channel with continuous activity (i.e. without pauses) is sufficient for measuring characteristics. The advantage of generating artificial voice is that it is more easily generated and have smaller variability than samples of real voice.

Gender (G): As discussed before, each language of 20 languages has two genders: female and male. In particular, each gender has 16 voice signals to variate the voice signal frequencies in the same gender type of one language. We believe that there are different human voices where the male voices usually have high frequencies and powers that the female ones. Therefore, gender type has to be a crucial feature in estimating the voice quality in VoIP networks, which is substantially involved in the experiments.

Burst ratio (B_{ratio}): Burst ratio metric [24] is considered one of the important parameters that have a major effect on voice quality. The metric is defined as a measure of burstiness in the packet-based network and is also defined as a ratio of the average length of observed bursts in a packet arrival sequences over the average length of bursts expected for a random loss in the packet-based network. The burst length refers to the number of packets in a single loss bursts. The burst ratio can be calculated by the following equation:

$$\text{Burst\_Ratio} = \frac{\text{MBL}_B}{\text{MBL}_R} = \frac{\left(\frac{\sum_{b=1}^{\text{Number}-\text{of}-\text{Burst}} \text{Busrt}-\text{length}_b}{\text{Number}-\text{of}-\text{Burst}}\right)}{\left(\frac{1}{1-L}\right)} \tag{3}$$

where

MBL_B	is the observed mean burst length of the loss sequence .
MBL_R	is the mean burst length of random packet loss .
$L = t/T$	is the the proportion of lost packets that is also a lost rate .
t	is the number of packets that are lost at the receiving end .
T	is the sending packets in a given period of time .

It has long been observed that loss on the packet-based network is bursty; which means that the loss of a packet depends on lost previous packets. This study used equation provided in ITU-T Recommendation G.107 [26] to compute the burst ratio according to the p and q values as follows:

$$\text{Burst\_Ratio} = \frac{1}{(p+q)} \tag{4}$$

Packet loss parameter uses burst ratio percentage to compute the p and q values in order to simulate degradation for a transmitted voice over a packet-based network. The proposed simulator takes packet loss and burst ratio percentage as input parameters for previous equations to compute p and q parameters to be used for inserting loss state in random positions of the transmitted voice stream.

Codec (C): The transmitted signal needs to be encoded in IP-network in order to reduce the quantity of data that transmitted over the network, in other words, reduce the channel bandwidth by compressing the data packets. In particular, En**Co**ding and **dec**oding (Codec) processes are used in data communications, networking, and storage as steps in transmitting voice over VoIP networks. The process of converting a sequence of characters (letters, numbers, punctuation, and certain symbols) into specially compressed formats for efficient transmission or storage is called encoding. In contrast, decoding is the process of converting an encoded format back into the original sequence of characters. Here, the two processes are utilised as features in the training set.

3.3 Classification techniques: building NIML approach

The NIML method takes the training set as input and uses machine learning techniques to extract patterns as output models. These models can be utilised for classifying future, unseen or unknown features. In particular, machine learning techniques extract the best hypothesis from a space set of hypotheses generated during the training phase. Each classifier has its own mathematical theory to map the input instances into output labels composed into a specific model or pattern.

In machine learning, there are classifiers categorised under the same theory but with small differences in technical issues. In this paper, we examine a set of classification learning techniques employed from different families. We use the most effective and accurate classifiers from each family. The selected classifiers in terms of family perspective include the following [28].

Bayes: It is a direct method that estimates the probability of the best hypothesis based on prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself. Then, Bayes methods build a rule-based or network model that reflects the probability of the best hypothesis. We use two classifiers which are NaiveBayes (NB), BayesNet (BN).

Functions: Each classifier aims to build a model of function (or hypothesis) where the input domain (i.e. features) is mapped into an output of range (i.e. labels). There are many forms of mapping such as the neural network (i.e. using feed-forward and back-propagation techniques for updating the network weights to reduce the error of loss function), logistics (regression), support vector machine (i.e. hyper-planes), and so on. Here, we use MultilayerPerceptron (MLP), SimpleLogistic (SL) and SMO.

Lazy: Simply keep the training data saved and only when it observes a test instance start generalisation to classify the instance based on its similarity to the saved training instances, in contrast with eager learners that construct a classifier as a ready learned model before observing testing instances to classify future instances such as Bayes, functions, meta, and trees families. We use three best algorithms that are IBk, KStar, and LWL.

Meta: The idea is to learn a set of classifiers (experts) from the training set. Each classifier is built by randomly selecting a subset of the training set (i.e. samples) with a replacement. Then, the ensemble classifier combines these weak learners for predicting future unseen test instances using voting or average methods. The ensemble classifier is constructed by two combining approaches as an independent (as voting or average) and a weighted sum of classifiers. The weighted sum ensemble classifier is built as weighted weak learner where a new classifier is learned from the incorrect classification of previous weak classifiers. Three best techniques in this family were used; AdaBoost (AB), Bagging (B), and LogitBoost (LB).

Trees: Each classifier in this family is built as a form of trees where each node represents the selected best attribute while the arcs represent the values of that attributes. Three techniques were examined which are Decision Tree (J48), LMT, RandomForest (RF).

3.4 Evaluation phase

For evaluation, particularly in the training phase, we repeat the experiments ten times for each classifier along with ten-fold cross-validation method. We also use a set of evaluation metrics, which are precision, recall, and receiver operating characteristic (ROC) area [29]. Precision (also called positive predictive value) is the percentage of relevant instances among the retrieved ones, while recall (also known as sensitivity) is the percentage of retrieved and relevant instances over the total amount of relevant ones. The ROC area (or curve) [30] is used to see how well your classifier can separate positive and negative instances and to identify the best threshold for separating them. In particular, the ROC curve is a graph plot that relates the true positive rate (sensitivity or recall) in function of the false positive rate for different threshold points of a parameter. Hence, each point on the ROC curve represents a recall/false positive rate pair corresponding to a particular decision threshold.

4 Experimental settings and results

In order to evaluate the proposed NIML prediction models, a simulation is built to emerge the VoIP processes as shown in Fig. 2. This section discusses the parameters values that are used in the simulation and presents the experimental results. The experimental results also include discussion and conclusion for answering the research questions in the Introduction.

4.1 Simulation settings

The NIML approach uses a training set of features and labels. We use the features that are discussed in Section 3. Each feature has a set of values that relate to MOS values as dependent metric labels. The parametric settings for each feature are discussed as processes as follows.

Encoding process: Two encoding methods were used in the simulation; which are ITU-T recommendation G.723.1 of 6.3 kbps [31] and ITU-T recommendation G.729 [32]. The ITU-T recommendation G.723.1 specifies a coded representation to be used for compressing the speech or other audio signal components of multimedia services at a very low bit rate. This coder has two-bit rates associated with it; which are 5.3 and 6.3 kbps. The higher bit rate has a high quality; the reason it is used by our experiments. The lower bit rate gives good quality and provides system designers with additional flexibility. Both rates are a mandatory part of the encoder and decoder. It is possible to switch between the two rates at any 30 ms frame boundary. An option for variable rate operation using discontinuous transmission and noise fill during non-speech intervals is also possible. This coder was optimised to represent speech with a high quality at the above rates using a limited amount of complexity. Music and other audio signals are not represented as faithfully as speech but can be compressed and decompressed using this coder. This coder encodes speech or other audio signals in 30 ms frames. In addition, there is a look ahead of 7.5 ms, resulting in a total algorithmic delay of 37.5 ms. All

Table 1 Frame information for G.729 and G.723.1

Codec	Algorithm	Bit rates, kb/s	Frame length, ms	Look-ahead, ms	Algorithmic delay, ms
G.729	CS-ACELP	8	10	5	15
G.723.1	MP-MLQ/ ACELP	5.3/6.3	30	7.5	37.5

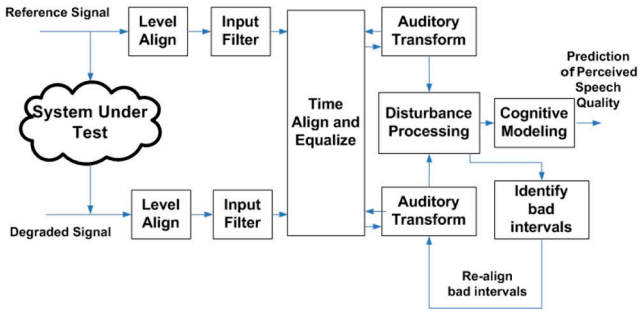

Fig. 4 *PESQ model [19]*

additional delays in the implementation and operation of this coder are due to:

1. actual time spent processing the data in the encoder and decoder;
2. transmission time on the communication link;
3. additional buffering delay for the multiplexing protocol.

ITU-T recommendation G.729 contains the description of an algorithm for the coding of speech signals at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP). This coder is designed to operate with a digital signal obtained by first performing telephone bandwidth filtering [33] of the analogue input signal, then sampling it at 8000 Hz, followed by conversion to 16-bit linear PCM for the input to the encoder. The output of the decoder should be converted back to an analogue signal by similar means. Other input/output characteristics, such as those specified by [33] for 64 kbit/s PCM data, should be converted to 16-bit linear PCM before encoding, or from 16-bit linear PCM to the appropriate format after decoding. The bitstream from the encoder to the decoder is defined within this recommendation.

The major differences between the two codecs lie in the excitation signals, the partitioning of the excitation space (the algebraic codebook), delay, and the way in which the coefficients of the filter are represented. The frame information for G.729 and G.723.1 is shown in Table 1. The delay induced at the encoder is referred to as an algorithmic delay.

The two codecs also have voice activity detection and silence suppression processing. The frames are classified as a normal speech frame, silence insertion description frame and null frame (non-transmitted frame).

Finally, these two encoding and decoding methods consequently result in a compressed signal to be an input for packet loss simulator for the next process.

Simulating packet loss: Simulating packet loss requires updating a set of bits in encoded packets using the 2-state Markov model of two probabilities values; p and q. The p and q values should be calculated from provided packet loss and burst ratio probabilities as depicted in Equations (2) and (4). Then, p and q values are used to add zeros in specific places in the encoded data stream as packet loss state using a specific threshold; which is the packet loss percentage. Hence, in order to compute the two probabilities, we use packet loss of 50 values from 1 to 50 percentages (i.e. increased by 1) and five values of 1 to 5 percentages for the burst ratio.

Decoding process: As known the received degraded signal at the end-side needs to be decompressed to retrieve the required original signal to play out at receiver buffer for listening, so the same two coders that were used for coding have an opposite process of encoding process that is a decoding process, hence the two decoders [31, 32] were used for decompression. The decompressed signal is a result signal that is utilised by the user at the receiver end. This decompressed or degraded signal is used as input along with the original signal to compute the voice quality as an objective estimation in the next step.

Estimating the voice quality using PESQ method: The PESQ is a standard method in [18, 19], which is used to compute the voice quality as MOS value. The method needs two signals as input to make the comparison and compute the quality of degraded (decompressed) signal, which are the original voice and the degraded (decompressed) signals where the output is MOS_{PESQ} value. This value is converted into MOS_{LQO} value. In more detail, the PESQ algorithm involves the following processing stages. First, the model aligns both the reference signal and the degraded signal to the same constant power level that corresponds to the normal listening level used in subjective tests to process a series of delay between two signals to determine the start and end points. Based on the set of delays that are found, PESQ compares the original (input) signal with the aligned degraded output of the device under test using a perceptual model. Then the signals are filtered using an fast Fourier transform based input filter to model and compensate or mimic for the filtering that takes place in the telephone handset and in the network. Both signals are aligned in time and then processed through an auditory transform similar to that used in PSQM and PAMS. The structure of the PESQ model is illustrated in Fig. 4.

The final PESQ score is a linear combination of the average symmetric disturbance value and the average asymmetric disturbance value, computed using the following formula [18, 19]:

$$\mathrm{MOS}_{\mathrm{PESQ}} = 4.5 - 0.1 * d_{\mathrm{SYM}} - 0.0309 * d_{\mathrm{ASYM}} \tag{5}$$

where $\mathrm{MOS}_{\mathrm{PESQ}}$ represents the P.862 PESQ MOS, d_{SYM} is the average symmetric disturbance value and d_{ASYM} is the average asymmetric disturbance value. The range of the PESQ score is between -0.5 and 4.5 as opposed to the ACR listening quality MOS which is on a 1–5 scale.

Converting MOS_{PESQ} to MOS_{LQO}: The PESQ values (i.e. $\mathrm{MOS}_{\mathrm{PESQ}}$) should be converted to a $\mathrm{MOS}_{\mathrm{LQO}}$ value, which is the quality estimation for a prospective user. As $\mathrm{MOS}_{\mathrm{PESQ}}$, that is defined in P.862, is calibrated against an essentially arbitrary objective distortion scale, it is not designed to be exactly the same scale as MOS. It is therefore desirable to provide an objective listening quality score from the P.862 that allows a linear comparison with subjective MOS. In order to achieve that and to align the PESQ MOS with the new MOS terminology as defined in [34], ITU-T published their recommendation [35]. This recommendation defines a mapping function and its performance for a single mapping from raw P.862 PESQ MOS scores to PESQ MOS-LQO. The mapping function is defined by [35]

$$z = 0.999 + \frac{4.999 - 0.999}{1 + \mathrm{e}^{-1.4945y + 4.6607}} \tag{6}$$

where

y is the P.862 PESQ score z is the corresponding PESQ P.862 MOS_LQO score

For normal subjective test material, the PESQ output range should be a listening quality MOS-like score between 1.0 (bad) and 4.5 (no distortion) [36]. This work follows a modified scale of the ACR test that uses five quality scores as bad (1), poor (2), fair (3), good (4), and excellent (5) [6].

The five processes were repeated 20,000 times to produce 20,000 records, which are a combination of the parameters as 50 packet loss values, 5 burst ratio values, 20 language, male and female and 2 codecs. Each value of the parameter along with $\mathrm{MOS}_{\mathrm{LQO}}$ value acts as a record to the training set. Table 2 summarises the parameters setting that were used in the simulation.

Table 2 Parameters settings of the experiment

Parameter	Value
packet loss (P_{loss})	{1,2,…, 50}
language (L)	20 languages {American English, British English, Arabic, …}
gender (G)	{Female, Male}
burst ratio (B_{ratio})	{1, 2, 3, 4, 5}
codec (C)	{G.723.1, G.729}

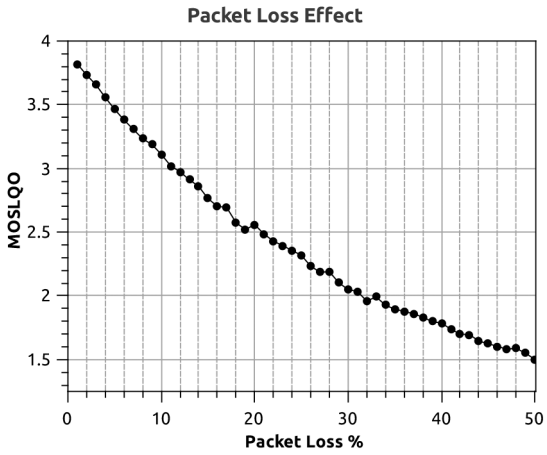

Fig. 5 *Packet loss percentage effect on MOS_{LQO} using G.723.1 codec*

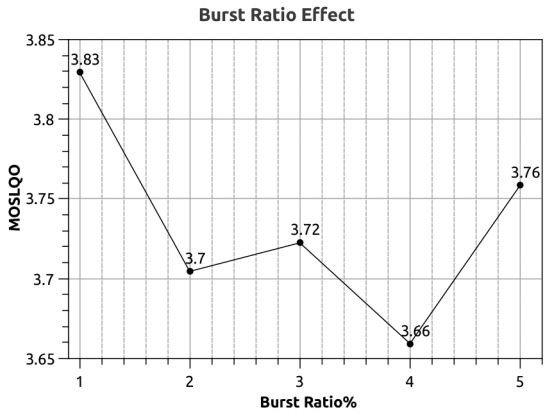

Fig. 6 *Burst ratio percentage effect on MOS_{LQO} using G.723.1 codec*

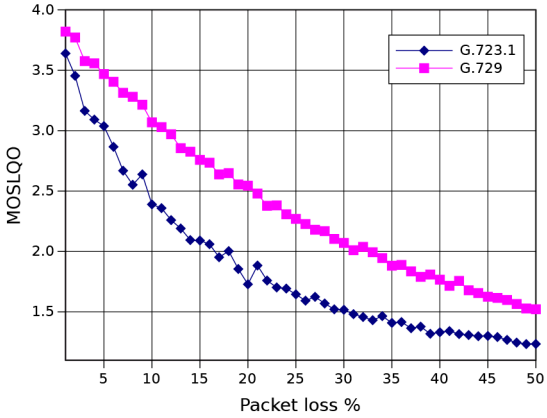

Fig. 7 *Codec effect on MOS_{LQO}*

4.2 Experimental results

This subsection presents the experimental results in two evaluation perspective. First, we study the effect of each parameter (i.e. five features) in the training sets on the voice quality to see how impact is the parameter as a feature on building the classifier models. Second, we present the evaluation of classifiers algorithms on the voice quality and a discussion includes the comparison of such algorithms.

Features effect and analysis: The features have impacts on the voice quality using their variation of parameter values. Here, we present the influences of each feature independently on the voice quality by initially setting the other features to fixed values.

To begin with, packet loss naturally has incremental decreasing effects with respect to their parameter values as shown in Fig. 5. The figure presents the percentage of packet loss on the x-axis, while MOS_{LQO} is presented at the y-axis. As shown in few places on x-axis the MOS_{LQO} values are increased due to the randomness in using 2-state Markov model for selecting the packet to remove it in the simulation. This randomness has an effect on the important parts in the original signal when it loses it.

As discussed, the loss in packet-based networks is burst in its nature. Fig. 6 shows the effect of increasing burst ratio to the voice quality, which in turn reflects losing packets in sequences. The consecutive packet lost (i.e. burst ratio) causes a problem when estimating the voice quality using PESQ algorithm. That is, the algorithm aligns the original and degraded signal to compare the signals, whereas turning the alignment with sequential packet loss rises a critical problem. Fig. 6 ensures the argument except at a case at burst ratio 5 due to an experimental issue where the effect is still in percentages.

In the encoding process, the algorithms split the signal into a set of bits using specific bits rate and frame length for creating the transmitted packets. Each encoding approach has a set of characteristics such as bit rate, frame length, and delay. If the encoding process uses an algorithm in high bit rate and frame length, this has an effect on the transmitted voice quality at the receiver end. Fig. 7 shows that the G.729 algorithm has high MOS_{LQO} values on voice quality in comparison with the G.723.1 algorithm by increasing the packet loss percentages, due to the low bit rate and large frame length using the G.723.1 codec. The results in Fig. 7 ensures the discussion about the parameters in Table 1 that G.729 technique outperforms the G.723.1 approach.

Several languages are available that might be used in VoIP and have an effect on transmitted voice quality. Hence, Fig. 8 shows the impact of using 20 languages on voice quality assessment. The x-axis presents the languages that are used on VoIP, while the y-axis represents the MOS_{LQO} corresponding values. The figure clarifies that diverse languages have a different effect on voice quality, which means that each distinct language has a power in its characters. Since the sounds in language determine its power and energy that involved in the frequency of the signal. The MOS_{LQO} values in Fig. 8 range between 3.25 and 3.82 on average of 3.5.

There is a distinct influence of using gender as a parameter on voice quality [37]. As shown in Fig. 9, the male gender has apparently a high impact on quality of transmitted voice using 50% of packet loss where the difference MOS_{LQO} values range between 0.2 (i.e. 1.3 and 1.5 at 50%) and 0.5 (i.e. 3.8 and 3.2 at 1%). This ensures different frequencies in using vowels between men and women.

In summary, these experimental results and study ensure the research question **RQ1** that each feature (or parameter) has a potential impact on voice quality in VoIP networks.

Evaluation results learning techniques: The features in training set independently manifest their influences on transmitted voice. The core part of the NIML method is the classification techniques for voice quality assessment. Hence, it is fundamentally important to explain how to use these techniques in predicting the quality of future degraded signals given their features. In order to evaluate the techniques, as discussed before we repeat the experiments ten times on each 10-fold cross-validation method. The evaluation metrics are used to potentially ensure the robustness of linear methods, where the metrics are precision, recall, and ROC area.

Table 3 shows the results of using 14 classifiers on a training set of 20,000 instances. The table differentiates each family with its classifiers. Principally, the NB classifier has better performance in precision value and less values in recall and ROC area compared to the BN classifier under Bayes family. This ensures the performance of the probability network of BN classifier. The MLP classifier has

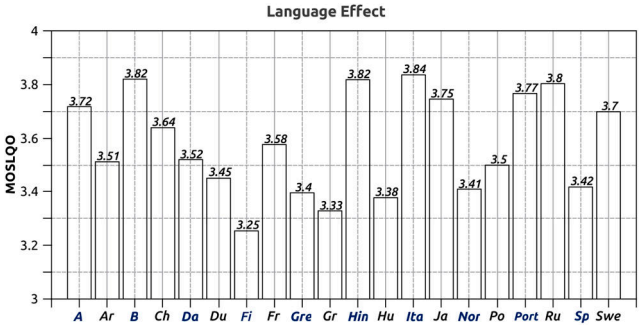

Fig. 8 *Languages effect on MOS_{LQO} using G.723.1 codec*

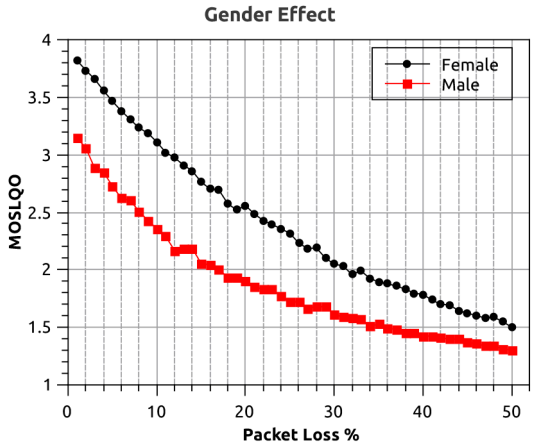

Fig. 9 *Gender effect on MOS_{LQO} using G.723.1 codec*

Table 3 Evaluation results on families of different classifiers

Family	Method	Precision	Recall	ROC area
Bayes	NB	**0.847**	0.841	0.952
	BN	0.846	**0.851**	**0.955**
Functions	MLP	**0.871**	**0.875**	**0.964**
	SL	0.862	0.864	0.96
	SMO	0.861	0.864	0.898
Lazy	IBk	***0.941***	***0.942***	0.968
	KStar	0.895	0.896	**0.982**
	LWL	0.747	0.809	0.933
Meta	B	***0.941***	***0.942***	**0.991**
	LB	0.853	0.854	0.959
	AB	0.746	0.805	0.885
Trees	RF	***0.941***	***0.942***	***0.992***
	J48	0.934	0.934	0.975
	LMT	0.931	0.931	0.982

The bold values are the best values on each family while the bold italic represents the best values for each metric (precision, recall, roc area).

a superior results over other classifiers of the same family on three evaluation metrics. The IBk classifier beats the KStar and LWL classifiers in precision and recall metrics and comparable performance with ROC area to the KStar classifier. In Meta and Trees families, the B and RF classifiers almost outperform the other classifiers on the same family with evaluation metrics precision (0.941), recall (0.942), and Roc area (0.99), respectively. The overall results across families, the IBk, B, and RF classifiers have superior performance over the other classifiers with apparently best performance allocated to RF classifier of precision 0.941, recall 0.942, and ROC area 0.992. These results lead to the research question **RQ2** affirmatively.

5 Conclusion and future work

The NIML voice quality assessment approach uses a training set of features with corresponding labels to build a classifier model for future unseen instances. The features are packet loss, burst ratio, codec, language, and gender, while the labels are the MOS_LQO values estimated from the PESQ_MOS values between the original and degraded signals. In this work, we studied the impact of such features independently on transmitted voice quality by using their MOS_LQO values and fixing the values of the remaining features. The results showed that each feature independently has an effect and can be used as features in the training set, which answers the research question **RQ1** affirmatively. On the other hand, a set of classifiers of different families were used on the training set to build models. The models take the values of features as input and predict the MOS_LQO values as a label. We used 15 classifiers of different families and evaluated their performance using three evaluation metrics. The results potentially ensure the ability to use such models in predicting the quality of transmitted voice over VoIP networks. Three classifiers of different families showed a superior impact on the others models, which are IBk, B, and RF classifiers. Hence, this answered the research question **RQ2** of robustness of using such models.

In future work, it is important to investigate a set of other features related to the voice itself such as frequency, time, phase, amplitude, and so on. Hence, signal-based features are being the target for evaluating the voice quality. In addition, a new set of classifiers including modern ones is being to be studied along with thorough discussions on their performance using different evaluation metrics.

6 References

[1] Karapantazis, S., Pavlidou, F.-N.: 'Voip: a comprehensive survey on a promising technology', *Comput. Netw.*, 2009, **53**, (12), pp. 2050–2090, Available at http://www.sciencedirect.com/science/article/pii/S1389128609001200

[2] Sun, L.: 'Speech quality prediction for voice over internet protocol networks'. Technical report, University of Plymouth, 2004

[3] Barry, M.A., Tamgno, J.K., Lishou, C., *et al.*: 'Challenges of integrating a VoIP communication system on a VSAT network'. 2017 19th Int. Conf. on Advanced Communication Technology (ICACT), Bongpyeong, South Korea, 2017, pp. 275–281

[4] Al-Akhras, M., Zedan, H., John, R., *et al.*: 'Non-intrusive speech quality prediction in VoIP networks using a neural network approach', *Neurocomputing*, 2009, **72**, pp. 2595–2608

[5] Rango, F., Tropea, M., Fazio, P., *et al.*: 'Overview on VoIP: subjective and objective measurement methods', *Int. J. Comput. Sci. Netw. Secur.*, 2006, **6**, (1), pp. 140–153

[6] Mahdi, A.E., Picovici, D.: 'Advances in voice quality measurement in modern telecommunications', *Digit. Signal Process.*, 2009, **19**, pp. 79–103, Available at http://www.sciencedirect.com/science/article/pii/S1051200407001832

[7] Raja, A., Azad, R.M.A., Flanagan, C., *et al.*: 'VoIP speech quality estimation in a mixed context with genetic programming'. GECCO '08: Proc. of the 10th Annual Conf. on Genetic and Evolutionary Computation, Atlanta, GA, USA, 2008, pp. 1627–1634

[8] Raja, A., Flanagan, C.: 'Genetic Programming, chapter Real-Time, Non-intrusive Speech Quality Estimation: a Signal-Based Model, 2008, pp. 37–48

[9] Soloducha, M., Raake, A., Kettler, F., *et al.*: 'Testing conversational quality of voip with different terminals and degradations'. 2017 Ninth Int. Conf. on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, 2017, pp. 1–3

[10] Salama, H., Dunne, J., Galvin, J., *et al.*: 'System for monitoring conversational audio call quality'. US Patent 9,635,087, 25 April 2017

[11] ITU-T Recommendation P.800: 'Methods for subjective determination of transmission quality'. International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), Geneva, August 1996

[12] Quackenbush, S., Barnawell, T., Clements, M.: '*Objective measures of speech quality*' (Prentice-Hall, Englewood Cliffs, NJ, 1988)

[13] Wang, S., Sekey, A., Gersho, A.: 'An objective measure for predicting subjective quality of speech coders', *IEEE J. Sel. Areas Commun.*, 1992, **10**, (5), pp. 819–829

[14] Yang, W.: 'Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measure Based On Audible Distortion And Cognition Model'. PhD thesis, Philadelphia, PA, USA, (May 1999), chair-Yantorno, Robert

[15] Voran, S.D.: '*U.S. Dept. of Commerce, National Telecommunications and Information Administration*' (Boulder, Colo., 1998), https://catalogue.nla.gov.au/Record/4136660

[16] ITU-T Recommendation P.861: 'Objective quality measurement of telephone-band (300–3400 Hz) speech codecs'. International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), Geneva, February 1996

[17] Rix, W., Hollier, P.: 'The perceptual analysis measurement system for robust end-to-end speech quality assessment', *Acoust. Speech Signal Process.*, 2000, **3**, pp. 1515–1518

[18] ITU-T Recommendation P.862: 'Perceptual evaluation of speech quality (PESQ):An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs'. International

Telecommunication Union-Telecommunication Standardization Sector (ITU-T), Geneva, February 2001

[19] Rix, A.W., Beerends, J.G., Hollier, M.P*., et al.*: 'Perceptual evaluation of speech quality (PESQ): a new method for speech quality assessment of telephone networks and codecs'. 2001 IEEE Int. Conf. on Proc. of the Acoustics, Speech, and Signal Processing (ICASSP '01), Salt Lake City, UT, USA, 2001, pp. 749–752

[20] Carvalho, L., Mota, E., Aguiar, R*., et al.*: 'An e-model implementation for speech quality evaluation in VoIP systems'. 10th IEEE Symp. on Computers and Communications (ISCC'05), Murcia, Spain, 2005, pp. 933–938

[21] Gaoxiong, Y., Wei, Z.: 'The perceptual objective listening quality assessment algorithm in telecommunication: introduction of itu-t new metrics polqa'. 2012 1st IEEE Int. Conf. on Communications in China (ICCC), Beijing, China, 2012, pp. 351–355

[22] Qawaqneh, Z., Mallouh, A.A., Barkana, B.D.: 'Deep neural network framework and transformed mfccs for speaker's age and gender classification', *Knowl.-Based Syst.*, 2017, **115**, pp. 5–14

[23] Sharan, R.V., Moir, T.J.: 'Robust acoustic event classification using deep neural networks', *Inf. Sci.*, 2017, **396(C)**, pp. 24–32

[24] McGowan, J.W.: 'Burst Ratio: A Measure of Bursty Loss on Packet-Based Networks'. United States Patent, B2 (6,931,017), August 2005

[25] Kekre, S.H., Saxena, H.B., C.L: 'A two-state Markov model of speech in conversation and its application to computer communication systems', *Comput. Electr. Eng.*, 1977, **4**, (2), pp. 133–141

[26] ITU-T Recommendation G. 107: 'The E-model, a computational model for use in transmission planning', International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), Geneva, March 2000

[27] ITU-TRecommendation P.50: 'Objective measuring apparatus'. International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), Geneva, September 1999

[28] Witten, I.H., Frank, E., Hall, M.A.: '*Data mining: practical machine learning tools and techniques*' (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011, 3rd edn.)

[29] Davis, J., Goadrich, M.: 'The relationship between precision-recall and roc curves'. Proc. of the 23rd Int. Conf. on Machine Learning (ICML '06), Pittsburgh, Pennsylvania, USA, 2006, pp. 233–240

[30] Lusted, L.B.: 'Signal detectability and medical decision-making', *Science*, 1971, **171**, (3977), pp. 1217–1219

[31] ITU-T Recommendation G.723.1: 'Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s', International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), March 1996

[32] ITU-T Recommendation G. 729: 'Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)', International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), March 1996

[33] ITU-T Recommendation G. 712: 'Transmission Performance Characteristics of Puls Code Modulation (PCM) channels'. International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), 1996

[34] ITU Recommendation P.800.1: 'Terms and definitions related to quality of service and network performance including dependability'. International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), Geneva, 1994

[35] I.-T.R. P.862.1: 'Mapping function for transforming P.862 raw result scores to MOS-LQO'. International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), Geneva, 2003

[36] Rix, A.W.: 'Comparison between subjective listening quality and p. 862 pesq score'. Proc. Meas. Speech Qual. Net. (MESAQIN), Prague, Czech Republic, 2003, pp. 17–25

[37] Fernandes, V., Ferreira, A.: 'On the relevance of f0, jitter, shimmer and hnr acoustic parameters in forensic voice comparisons using gsm, voip and contemporaneous high-quality voice recordings'. Audio Engineering Society Conf.: 2017 AES Int. Conf. on Audio Forensics, Arlington VA, USA, 2017