Intelligent Learning Techniques applied to Quality Level in Voice over IP Communications

Demostenes Zegarra Rodriguez, Renata Lopes Rosa, and Graça Bressan
Department of Computer Science and Digital Systems
University of São Paulo, SP - Brazil
Email: demostenes@larc.usp.br, rrosa@usp.br, gbressan@larc.usp.br

***Abstract*—This paper presents a method for determining the quality of a Voice over IP communication using machine learning techniques. The solution proposed uses historical values of network parameters and communication quality in order to train the different learning algorithms. After that, these algorithms are able to find the quality of the Voice over IP communication based on network parameters of a specific period of time. Intelligent and other machine learning algorithms take as input a baseline file that contains some values of network parameters and voice coding, associating an index quality for each scenario according to ITU-T Recommendation G.107. The tests were performed in an emulated network environment, totally isolated and controlled with real traffic of voice and realistic IP network parameters. The quality ratings obtained for the learning algorithms in all the scenarios were corroborated with the results of the algorithm of ITU-T Recommendation P.862. The results show the reliability of the four learning algorithms used on the tests: Decision Trees (J.48), Neural Networks (Multilayer Perceptron), Sequential Minimal Optimization (SMO) and Bayesian Networks (Naive). The highest value of reliability for determining the quality of the Voice over IP communications was 0.98 with the use of the Decision Trees Algorithm. These results demonstrate the validity of the method proposed.**

Keywords-QoS; VoIP; Machine Learning; MOS; E-Model; PESQ.

I. INTRODUCTION

The quality of a Voice over IP (VoIP) communication does not have the quality levels of the conventional circuit-switched telephony; thus, users who do not have an acceptable user experience with VoIP calls continue using traditional telephony. For this reason, the study of methods for evaluating quality of a VoIP communication is very important because it allows network resources to be reallocated to improve communication quality [1].

Initially, the determination of the quality of a VoIP communication was conducted by subjective tests, resulting in a quality score called MOS (Mean Opinion Score); ITU-T Recommendation P.800 [2] describes the requirements and methodology followed in these tests. Later, some objective methods were employed, such as ITU-T Recommendation P.862 [3] or PESQ (Perceptual Evaluation of Speech Quality), which determine an index named MOS-LQO (MOS-Listening Quality Objective), which is the result of the comparison of the original speech or reference signal and the degraded speech signal. Also, nonintrusive methods were developed, such as ITU-T P.563 [4], the algorithms of which do not need a voice signal reference.

Other metrics of voice quality [5][7] do not consider the voice signal, for determining the quality index models based on network parameters such as the E-Model [5] are employed, used in network planning and to configure the rate of VoIP communications.

Machine Learning algorithms have been used to determine the quality of multimedia services, thus trying to ensure a better quality of service [8][9]. For evaluating voice quality in VoIP services, [10][13] show how neural networks are used for monitoring this service, but other learning techniques are not studied and do not detail how the training file used was built. It is worth noting that, the E-Model is not sufficient to predict the voice Quality Level, because sometimes a parameter is missing and it is not possible to measure the QoS with the E-Model. Conversely, with machine learning, if one parameter is missing, it is possible to measure the QoS. In this context, this paper uses as a training file built based on ITU-T Recommendation G.107, better known as E-Model, considering some network scenarios that were extracted from real traffic. As a consequence, the results attained high levels of reliability.

The algorithms used in this study come from different approaches in the artificial intelligence sub-area devoted to the study of machine learning to predict the quality level of a service. These algorithms are: Decision Tree (J48 - C4.5 algorithm), Bayesian networks (Naive Bayes), Sequential Minimal Optimization (SMO) and Neural Networks (Multilayer Perceptron), and are used to determine the quality of a VoIP communication in a sample interval of 8 seconds. The reliability of the algorithms specified for this application was measured. Network training was performed using a 650-case file, prepared by the E-Model algorithm. Each line of the file contains the network parameters: transmission rate, delay and packet loss probability, and the value of voice quality index (MOS), which is the result of the E-Model algorithm.

This work considers the encoding rates of 64 kbps for ITUT G.711 [14] codec and rate of 8 kbps for ITU-T G.729 [15], respectively, and also considers the intrinsic values that these codecs have in the scenario with packet loss. The tests were performed in a scenario of IP network emulation, where a VoIP communication is established and different network parameters are programmed, in order to study the quality degradation of voice communication for each test scenario. For network

emulation purposes, a network emulator software was used. Thus, the parameters of packet loss probability and delay in an IP network were changed.

This article is divided as follows: Section II makes a theoretical revision of the machine learning algorithms used in this work. Section III deals with the voice quality assessment methodologies. Section IV presents the test scenario, the methodology followed for the tests and the parameters evaluated. Section V shows the experimental results and discussions, and finally, Section VI presents the conclusions and future work.

II. Algorithms used in the Determination of Voice Communication Quality

Artificial intelligence and machine learning appeared in the mid 1950s. The ability to learn is the main characteristic of artificial intelligence. This section presents the different algorithms used for training the IP network to determine the quality of VoIP communications.

The proposal of this work is to study how the artificial intelligence can help to find the adequate level of quality for a voice communication, because in [16] the artificial intelligence is used to find key metric for QoS in VoIP applications, but studies only the packet loss. In [17] the packet loss and jitter are considered, but only some artificial algorithms are studied.

A. Decision Tree Classification

Decision Trees [18] are tools that can be used for giving the agent the ability for both learning and making decisions. The decision tree takes as input a situation described by a set of attributes and returns a decision, which is predicted by the value of the input attribute. The input can be both discrete or continuous values. Only discrete values are used herein. The learning of discrete values is called classification.

To better understand the operation of a decision tree, it is considered the problem of choosing the correct QoS (Quality of Service) regarding to an IP network parameter, for instance, the delay, as shown in Figure 1.

B. Bayesian Classification

The Bayesian classification [19] algorithm has its name because it is based on Bayes Theorem probability. It is also known as Naive Bayes classifier or only Bayes algorithm. The algorithm aims to calculate the probability of an unknown sample belonging to each of the known classes. This type of prediction is called statistical classification; it is completely based on probabilities.

A feature of this algorithm is that it requires a data set previously classified. Based on this preliminary data set, also called training set, the algorithm takes as input a new unknown, i.e., which has no classification, and returns as output the most likely class for this sample according to probabilistic calculations.

The probability model for a classifier is a conditional model over a dependent class variable C with a small number of classes, dependent on several feature variables F_1 through F_n.

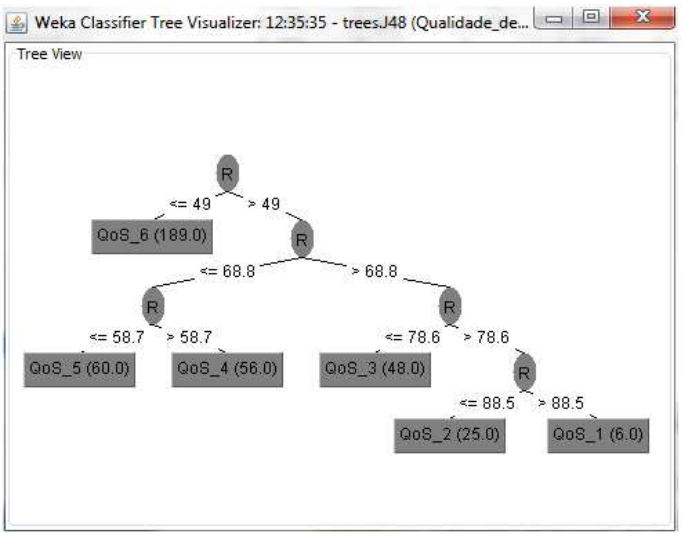

Figure 1: Tree algorithm classification used to determine the QoS, regarding the delay parameter.

C. Multilayer Perceptron

Rosenblatt [20] introduced the perceptron as a simple algorithm for neural network, capable of linearly classifying separable patterns. The operation of a perceptron (artificial neuron) shows that:

- The neuron is responsible for calculating the combination of inputs and weights, and then applies an activation function that determines the effective output of the neuron.
- The training is performed through the presentation of known inputs and outputs (supervised learning) and through adjusting the weights with specific algorithms.

Multilayer Perceptron Networks (MLP) is computationally more powerful than networks without hidden layers. MLP can handle data that are not linearly separable.

The processing performed by each neuron is defined by the combination of the processing performed by the previous neurons layer connected to it. From the first hidden layer to the output layer, implemented functions become increasingly complex. These functions define how the space-making division is made. There are several algorithms to train MLPs.

Among these, the most popular learning algorithm for training these networks is back propagation [21]. This is a supervised algorithm that uses the desired output for each input provided to adjust the parameters, called weights of the network. In addition, the adjustment weights use the backpropagation gradient method to define the corrections to be applied.

Figure 2 illustrates a perceptron network with three layers.

This network has an input layer (on the left) with three neurons, one hidden layer (in the middle) with three neurons and an output layer (on the right) with three neurons.

There is one neuron in the input layer for each predictor variable. In the case of categorical variables, $N - 1$ neurons

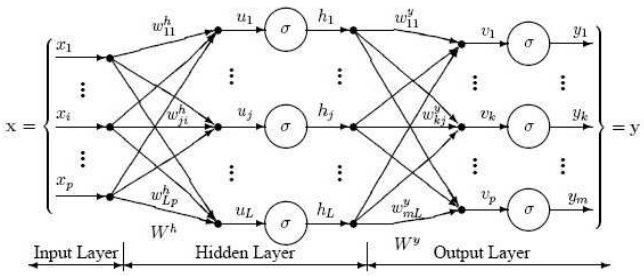

Figure 2: Perceptron network with three layers.

are used to represent the N categories of the variable.

- Input Layer: A vector of predictor variable values ($x_1 ... x_p$) is presented to the input layer. The input layer (or processing before the input layer) standardizes these values, so that the range of each variable is -1 to 1. The input layer distributes the values to each of the neurons in the hidden layer. In addition to the predictor variables, there is a constant input of 1.0, called the bias that is fed to each of the hidden layers; the bias is multiplied by a weight and added to the sum going into the neuron.
- Hidden Layer: Arriving at a neuron in the hidden layer, the value of each input neuron is multiplied by a weight ($w_j i$), and the resulting weighted values are added, producing a combined value u_j. The weighted sum (u_j) is fed into a transfer function, which generates a value h_j. The outputs from the hidden layer are distributed to the output layer.
- Output Layer: Arriving at a neuron in the output layer, the value from each hidden layer neuron is multiplied by a weight ($w_k j$), and the resulting weighted values are added, producing a combined value v_j. The weighted sum (v_j) is fed into a transfer function, which outputs a value y_k. The y values are the outputs of the network.

If a regression analysis is performed with a continuous target variable, then there is a single neuron in the output layer, and it generates a single y value. For classification problems with categorical target variables, there are N neurons in the output layer producing N values, one for each of the N categories of the target variable.

D. Sequential Minimal Optimization

It is an algorithm described in [22], in which a problem is decomposed successively into two subproblems, decreasing the number of vector operations necessary to resolve the problem. Training a support vector machine requires the solution of a very large quadratic programming (QP) optimization problem. Thus, SMO breaks this large QP problem into a series of smallest possible QP problems. At every step, SMO chooses two Lagrange multipliers to jointly optimize, finds the optimal values for these multipliers, and updates the SVM to reflect the new optimal values.

Two Lagrange multipliers can be done analytically, and this is the main advantage of SMO. Thus, numerical QP optimization is avoided. The inner loop of the algorithm can be expressed using a small code, rather than invoking an entire QP library routine. Also, more optimization sub-problems are solved in the course of the algorithm, each sub-problem is so fast that the overall QP problem is solved quickly. Therefore, SMO requires no extra matrix storage at all. Because no matrix algorithms are used in SMO, the probability to present numerical precision problems is low [22].

SMO is composed by an analytic method for solving for the two Lagrange multipliers, and a heuristic for choosing which multipliers to optimize.

III. Voice Quality Assessment Methodologies

The voice quality assessment is intended to give a score to a particular communication. These results are used to make improvements in transmission networks and, in general, for all the equipment involved in this process.

In this work, Recommendations ITU-T P.862 [3], ITU-T P.563 [4] and ITU-T G.107 [5] will be used in different network scenarios, in which different codecs will be used. The voice quality assessment methodologies and their classification will be shown in the following sub-sections.

A. Methodology Classification

1) Nonintrusive Method: In this type of method, a sample of the original signal communication is not necessary; the evaluation is only determined by the signal at the point the analysis is performed.

The nonintrusive method can be of two types:

- Objective, if one uses a tool (software) for the analysis and calculation of the quality score;
- Subjective, if a listener directly intervenes in the score regarding the quality of the results.

The recommendations below are examples of these methods:

- Objective methods: ITU-T G.107 (E-Model), ITU-T P.563.
- Subjective method: ITU-T P.800 [2].

2) Intrusive Methods: The intrusive method is one that requires a speech sample at the communication origin point, in order to compare with the destination point sample. As result of this comparison a voice quality index is given. Examples of such method are:

- Methods Objectives: Rec. ITU-T P.861 [23] and Rec. ITU-T P.862.

The Figure 3 is adapted from [2] and presents the difference between these two methods.

The following items are described in the ITU-T recommendations P.800, P.862, P.563 and G.107, and these recommendations are used in the experimental tests.

B. ITU-T Recommendation P.800

This recommendation provides a guide to conduct subjective tests of transmission quality in laboratories, and aims to indicate the methods considered appropriate for determining the

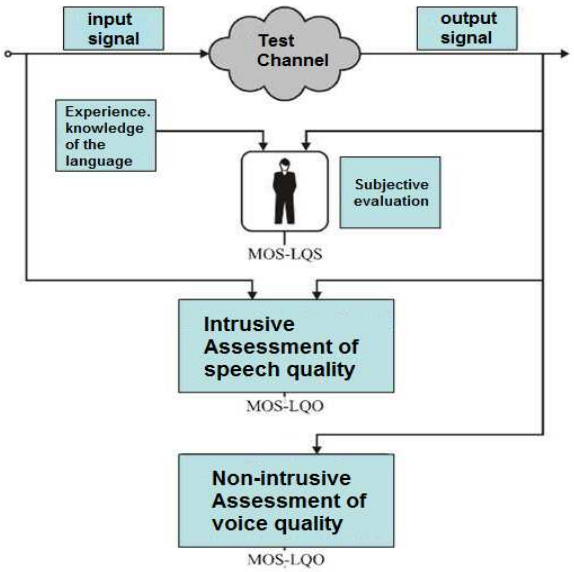

Figure 3: Difference between intrusive and non-intrusive methods [2].

degree of satisfaction by users using a voice communication service.

The recommended methods are:

a) Tests of opinion regarding a conversation.
The conversation tests aim to reproduce in a laboratory the actual conditions of a telephone service. For this, it is necessary to properly select the network conditions and participants, as well as conducting the tests appropriately.
Annex A of recommendation P.800 details the considerations to perform the tests in order to obtain reliable results.
These tests should be performed in a place of dimensions not less than 20 m^3 with a reverberation time smaller than 500 ms, with noise levels that approximate the characteristics of a library or hospital.
In the selection of test participants, the following conditions are required:

- Participants should not have participated directly in the studies of quality evaluation of voice communication services or similar tasks, such as the encoding voice signals.
- The participants should not have participated in any subjective test in the last six months and no conversation tests in the last twelve months.

The ITU-T recommends the following methodologies to be used in the tests:

- Opinion range of the conversation: 5 different scores can be used to assess the various quality categories. The opinion score presented in Table I is the most widely used. In these tests, the listener gives the adjective of the left column and the test lead makes the individual equivalence with the number in the right column. This procedure is valid for other types of tests that are presented in Tables I, II and III.
The arithmetic mean of any set of these scores is called mean opinion of score, which is represented by MOSc (Mean Opinion Score of conversation).
- Difficulty scale: This is a binary response obtained from each participant at the end of each conversation. The question to be performed is: Does the interlocutor or do you experience any difficulty in speaking or listening through the connection? Answer: Yes, No.
Assigning the following values: Yes = 1 and No = 0. The amount assessed (percentage of yes answers) is called the difficulty percentage and is expressed by the symbol: %D.

b) Listening tests - Index determination by Absolute Category Rating (ACR).
As this test does not achieve the same degree of realism as conversation tests, the considerations are less stringent. However, this requires a strict control of certain parameters such as: recording the source voice signal, an adequate calibration of the emitter system, the listening room should have the same conditions as those in a recording room. Regarding the source signals, it is essential to use more than one male voice and one female voice to reduce the risk, since the results depend on the peculiarities of the voices chosen.
The test participants (listeners) are chosen in a population that typically uses telephone services, with the following conditions: not having directly participated in jobs related to the assessment of transmission quality of telephone services, or similar tasks, not having participated in subjective tests for at least six months and listening tests in a year, and having no previous contact with the test sentences.
The opinion scales recommended by the ITU-T for this type of evaluation are:

- Listening-quality scale: the average magnitude of these scores is represented by the symbol MOS. The scores for each category in this scale are presented in Table I.
- Listening effort scale: the average magnitude of these scores is represented by the symbol MOSle. The scores for each category in this scale are presented in Table II.
- Audibility preference scale: the average magnitude of these scores is represented by the symbol MOSlp. The scores for each category in this scale are presented in Table III.
In the ACR method, Annex D from Recommendation ITU-T P.800 defines the DCR

method, which is basically a comparative evaluation between the high-quality original voice and the sample to be evaluated. This method is used when the degradation is small. The scale goes from inaudible to very poor quality. Therefore, it is useful to optimize the system, after ACR method determines that the degradation condition is within the range of acceptable quality.

TABLE I: Score for Conversational MOS and hearing tests

Signal quality	Score
Excelent	5
Good	4
Regular	3
Bad	2
Very bad	1

TABLE II: MOS score for listening effort test

Effort required to understand the meaning of the sentences	Score
Perfect listening, no effort	5
Some attention is necessary, no appreciable effort	4
Moderate effort	3
Considerable effort	2
Meaning incomprehensible, with greater effort	1

TABLE III: MOS score for listening preference test

Audibility Preference	Score
Much greater than preferred	5
Greater than preferred	4
Preferred	3
Smaller than preferred	2
Much smaller than preferred	1

C. ITU-T Recommendation G.107

This recommendation, better known as E-Model [24], is a mathematical and computational model that measures the effects of the network parameters over the voice quality transmitted. The E-Model is defined by the following equation:

$$R = R_o - I_s - I_d - I_e + A \quad (1)$$

In which:

- R: determination factor of the transmission rate that has a correspondence with the MOS score ITU-T P.800.
- R_o: signal to noise ratio.
- I_s: simultaneous degradation factor, which represents all the degradations that occurs simultaneously with the speech signal.
- I_d: quality degradation due to the delay.
- I_e: quality degradation deriving from the device (codec).
- A: improvement factor.

The default value of R_o is 93.2, which is obtained by setting all model variables with default values, for example, the parameter of quality degradation due to delay (I_d), and the parameter corresponding to the equipment degradation (I_e) do not consider the packet loss rate for this calculation. As a result, R_o reaches a high value closer to 100.

The parameter I_s is not considered in this work, since it describes conditions that are related to the signal, not depending on the transport network. The factor A has the value 0 [5] for cable networks, which matches with the emulation scenario used in this work.

The delay factor I_d is defined by (2):

$$I_d = I_{dle} + I_{dte} + I_{dd} \quad (2)$$

Parameters I_{dte} and I_{dle} correspond to delays due to echo, for the sender and receiver, respectively. These factors are not considered for the test scenario assumption of perfect echo suppression. The I_{dd} represents the delay (T_a) produced in both, the codec and the network, respectively. The network delay is set in the network emulator according to the type of test to be performed.

The I_{dd} is defined as:

For $T_a \leq 100ms$: $I_{dd} = I_d = 0$

For $T_a > 100ms$:

$$I_{dd} = I_d = 0.024d + 0.11(d - 177,3)P(d - 177,3) \quad (3)$$

With: $P(k)$ = 0, if $k < 0$, $P(k)$ = 1, if $k \geq 0$

With these considerations, the parameter R can be calculated as a function of the parameters that correspond to the value of R_o, the delay (I_d) and the factor corresponding to the codec (I_e).

$$R = R_o - I_d - I_e \quad (4)$$

The R factor is related to the MOS index, according to the following equation:

$$R = 3,026 \times M^3 - 25,314 \times M^2 + 87,06 \times M - 57,336 \quad (5)$$

The relation between R parameter and MOS index is presented in Figure 4 [5].

D. ITU-T Recommendation P.862

The ITU-T Recommendation P.862 [3], known as PESQ is an objective evaluation method that compares an original with a degraded voice signal, resulting from the passage of the voice signal through a communication system. The output is a PESQ quality index, which predicts the perception of quality that would be perceived by an assessor during a subjective listening test. The subjective quality perception is related to a score, called MOS (Mean Opinion Score) index; whereas, the PESQ algorithm estimates the index MOS-LQO and the range used by PESQ goes from 1 (poor) to 4.5 (excellent).

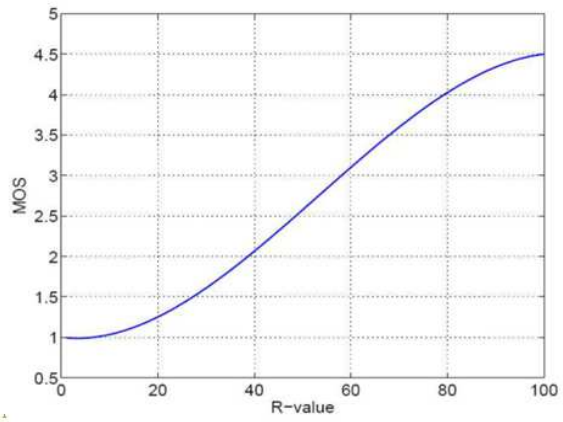

Figure 4: Relation between R parameter and MOS index.

PESQ only measures the effects of one-way speech distortion and noise on speech quality. The effects of delay, echo, and other impairments related to two-way interaction are not reflected in the PESQ scores. According to this recommendation, PESQ demonstrated acceptable accuracy in the following scenarios:

- Speech input levels to a codec.
- Transmission channel errors.
- Packet loss and packet loss concealment with CELP codecs.
- Transcodings.
- Effect of varying delay in listening only tests.
- Short-term and long-term time warping of audio signal.
- Coding Technologies: Waveform codecs, e.g., G.711; G.726; G.727; CELP and hybrid codecs .4 kbit/s, e.g., G.728, G.729, G.723.1; Other codecs: GSM-FR, GSMHR, GSM-EFR, GSM-AMR, CDMA-EVRC, TD-MAACELP and TDMA-VSELP.

The G.711 and G.729 codecs were used in the test scenarios, and the network suffers degradation due to packet loss and delay. The audio input was isolated to avoid voice impairments. Only delay is not included in the scenarios of recommendation P.862 tested. To make possible to include this factor in the test scenarios where the delay is present, the MOS index was converted to R index value using equation (5) to obtain the quality index in this type of scenarios.

a) Description of the algorithm used by the PESQ
The PESQ compares an original voice signal with a degraded voice signal, resulting from several types of degradation. The PESQ output is named as MOS-LQO.
In a first step, a series of delays between the original and degraded voice signals are computed, one for each time interval in which the delays are significantly different between the two signals. For each interval, the start and end points are calculated. The alignment algorithm is able to handle delay changes in both periods of silence and in speech.
Figure 5 shows the basic idea used in PESQ. The computational model compares the input and output of the device being tested, comparing the original and degraded output, this degradation is caused by delay or packet loss of the network.
Based on the set of delays encountered, the original and degraded signal already aligned are compared using a perceptual model, as illustrated in Figure 5. The idea of this process is to make both signals into a form of internal representation that is analogous to the psychophysical representation of the signal in the human auditory system, taking into account perceptual frequency and intensity. This is achieved at several stages: time alignment, intensity level alignment, time-frequency mapping, frequency scale transforming, and intensity range compression.

b) Variables that influence the PESQ performance
Tables IV, V and VI show a summary of the factors that must be considered to obtain valid results from the tests.
Table IV shows the test factors, coding technologies and applications in which PESQ has an acceptable performance.
Table V presents some variables in which the PESQ method has unreliable predictions.
Table VI shows factors, technologies and applications for which the PESQ was not evaluated.

TABLE IV: Scenarios in which PESQ algorithm has an acceptable performance

Tests Factors
Input levels of speech signal in a codec.
Errors in the transmission channel.
Packet loss and packet loss concealment with CELP codecs.
Multirate codecs.
Ambient Noise on the sending side.
Coding technology
Waveform codecs, for example: G.711; G.726; G.727.
CELP and hybrid codecs with rate greater than 4 kbit/s,
for example: G.728, G.729, G.723.1.
Other codecs applications: GSM-FR, GSM-HR, GSM-EFR, GSM-AMR,
CDMA-EVRC, TDMA ACELP, TDMA-VSELP.
Aplications
Evaluation of codecs.
Selection of codecs.
Evidence in emulated networks and prototype networks.

TABLE V: Test scenarios unsuitable for PESQ algorithm

Predictions that have no significance
Loss of loudness.
Delay, delay effect in conversational tests.
Eco perceived by the speaker.
In the process of coding with cut voice continuous sections where this period extracted represents more than 25% of the total period of voice activity.
Behavior quality of two-way communications.

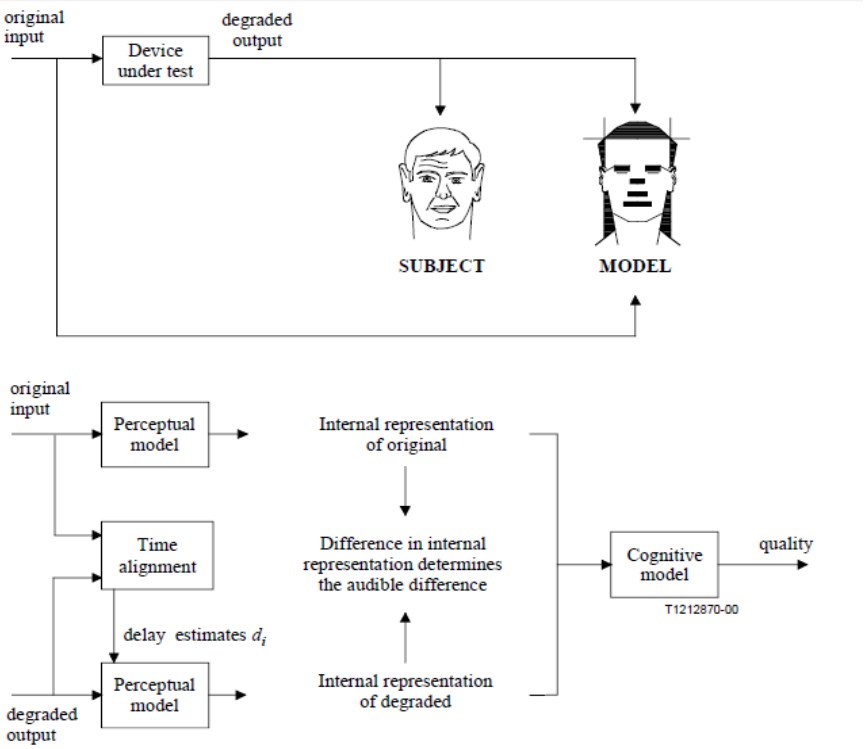

Figure 5: Algorithm used by PESQ [2].

TABLE VI: Test scenarios for which the PESQ has not been evaluated

Testing Factors
Replacement of speaking sessions for silence.
Amplitude saturation of the voice signal.
Effects of dependence with respect to the speaker.
Simultaneous speakers.
Music as codec input.
Eco to the listener and the effects of echo and noise compensators.
Coding technology
CELP and hybrids codecs with rates lower than 4 kbit/s.
MPEG4.
Aplications
Proofs of acoustic terminals through the head and back simulator.

E. ITU-T Recommendation P.563

Algorithm P.563 is applicable to speech quality evaluation without requiring a reference signal. For this reason, it is recommended to a non-intrusive assessment of voice quality and for the supervision of a real network based on only one extreme. This recommendation is not limited to end-to-end measurements, and can also be used at any point in the transmission chain. The score, thus calculated, is comparable to the quality perceived by a human listening to the signal at the test point. This MOS index evaluation is termed MOS-LQO.

This recommendation should be used to assess the speech quality in telephony applications of 3.1 kHz bandwidth. It should be emphasized that algorithm P.563 does not provide a complete assessment of the transmission quality. Thus, it only measures the effects of unidirectional voice distortion and noise in the voice quality, in the same way as a listening test, which assesses the quality in ACR scale. This means that the effects of loss of sound, delay, echo of the speaker, and others that affect the two-way interactions did not influence the P.563 scores. It is worth nothing that this algorithm is designed exclusively for the evaluation of human voice, and can not be applied for music or, generally, other non-vocal audio signals.

The digitized voice signal must meet the following requirements:

- Sampling frequency: 8000 Hz;
- Linear PCM amplitude resolution of 16 bits;
- Minimum duration of 3.0 s and a maximum of 20 s;
- Minimum ratio of vocal activity of 25 % and a maximum of 75 %.

IV. Scenario and Test Methodology

The tests were conducted in the emulation of IP network shown in Figure 6, which consists of three computers; two of them (PC-A and PC-C) establish point to point VoIP communication, and the third (PC-B) emulates the transmission channel, which programmed different degradations of network, such as packet loss and delay. The audio inputs used belong to the set of validation tests that are included in ITU-T P.862, the algorithm of the same recommendation (PESQ) is used to assess the level of quality of voice signal in reception.

The second quality index is determined by machine learning algorithm; these algorithms are decision trees, neural networks, SMO and Bayesian networks. The first step is to build the algorithm used in machine learning or training file, which was built considering 650 scenarios, where QoS was obtained from the E-model algorithm. Each test scenario is represented in the training file as a line, which will be presented later in more detail. In the emulator network, different scenarios are programmed, with different parameter values of probability of packet loss and delay inserted in the training file in order to better validate the reliability of results. The sample period for assessing the quality of VoIP communications is 8 seconds. This time was chosen, in addition to the original size of the audio file, due to the number of packets sent by the transmitter (PC-A) during this time period. Thus, considering an encoding rate of 64 kbps, 400 packets of 160 bytes are sent; for a shorter time, the number of packets sent is smaller and, therefore, the percentage of packet loss has a lower resolution.

The software and tools, in each PC used to build the test scenario are the following:

- PC-A, PC-C: clients using softphone MyPhone 0.2b10 [25], packet analyser Wireshark [26] and software to record audio, a .wav file,VRS Recording System [25].
- PC-B: router using the network emulator NETEM [28], to simulate packet loss, delay, jitter and bandwidth; software ITU-T P.862 was used to find the MOS index for voice quality.

The sound transmitted was generated by a player of a .wav file that is connected to the microphone input of PC1 by an audio cable. As mentioned previously, this file has a duration of 8 seconds and was sampled at 8 kHz and 16 bits.

The methodology followed in the tests to get the value using the tool PESQ MOS is as follows:

- Initially, it starts a communication between the PC-A and PC-C by softphones installed in the PCs, and lets you choose the voice codec used to each VoIP call, made from computer to computer.
- For each scenario, the network emulator is configured with parameters required to perform the test.
- The player transmits the audio (arq-orig.wav) to PC-A where this sound is recorded (arq-orig2.wav), as the call is active, and the voice is transmitted to the PC-C through the PC-B.
- While data are transmitted, the program Wireshark running in PC-A and PC-C saves network information, such as the signaling messages for establishing, maintaining and finalizing calls, messages from the RTP, the average size of the package (bytes), average number of packets transferred per second and average bandwidth.
- In the PC-C, the audio received is recorded in a file (arq-deg.wav). This file and the original file are compared by software PESQ, which runs on PC-B, resulting in a MOS-LQO score.

A. Implementation of the input file for training algorithms

As aforementioned, the preparation of an initial database of network parameters and the quality score for each scenario was performed considering the ITU-T Recommendation G.107. This mathematical model provides an R value of quality ranging from 0 to 93.2, in which the highest value corresponds to a higher quality. The different scenarios were done leaving the default parameters fixed and varying the following parameters:

- The Mean One-Way Delay.
- The Packet-loss Probability.
- The type of codec is related to the parameters: encoding rate, I_e (Equipment Impairment Factor) and Bpl (Packet-loss Robustness Factor).

The values of I_e and Bpl are dependent on the vocoder used. Table VII presents the values of these codec parameteres that have been tested by the ITU-T recommendation G.107. In our test scenarios, codecs G.711 and G.729 were employed.

TABLE VII: Values of I_e and Bpl for voice codecs

Codec	I_e	Bpl
G.723.1+VAD	15	16.12
G.729A+VAD	11	19
GSM-EFR	5	10.03
G.711	0	4.3
G.711+PLC	0	25.14

The parameters described in Table VIII were used to obtain the value of quality index R.

TABLE VIII: Parameters of the E-Model Algorithm

Parameter	Default Value	Units
Noise Referred to at 0 dBr point - Nc	-70	dBm
Noise Floor - Nfor	-64	dBm
Room Noise (Receive) - Pr	35	dB
Send Loudness Rating - SLR	8	dB
D-factor (Receive) - Dr	3	-
Listener's Sidetone Rating - LSTR	21	dB
D-factor (Send)	3	-
Mean One-Way Delay - T	100	ms
Absolute Delay from (S) to (R) - *T*a	100	ms
Round-Trip Delay - *T*r	200	ms
Weighted Echo Path Loss - WEPL	110	dB
Quantizing Distortion Units - QDU	1	-
Equipment Impairment Factor - Ie	0	-
Packet-loss Robustness Factor - Bpl	1	-
Packet-loss Probability - Ppl	1	%
Expectation Factor - A	0	-

As a result, a file with 650 lines was obtained; for better understanding, Table IX presents the 10 first cases or network scenarios.

TABLE IX: Values of I_e and Bpl for voice codecs

Rate (kbps)	Delay (ms)	Bpl (%)	Ie	R (Value)
64	0	4.3	0	93.2
64	50	4.3	0	91.8
64	100	4.3	0	90.7
64	150	4.3	0	89.5
64	200	4.3	0	85.8
64	250	4.3	0	79.2
64	300	4.3	0	72.5
64	350	4.3	0	67
64	400	4.3	0	62.2
64	450	4.3	0	58.2

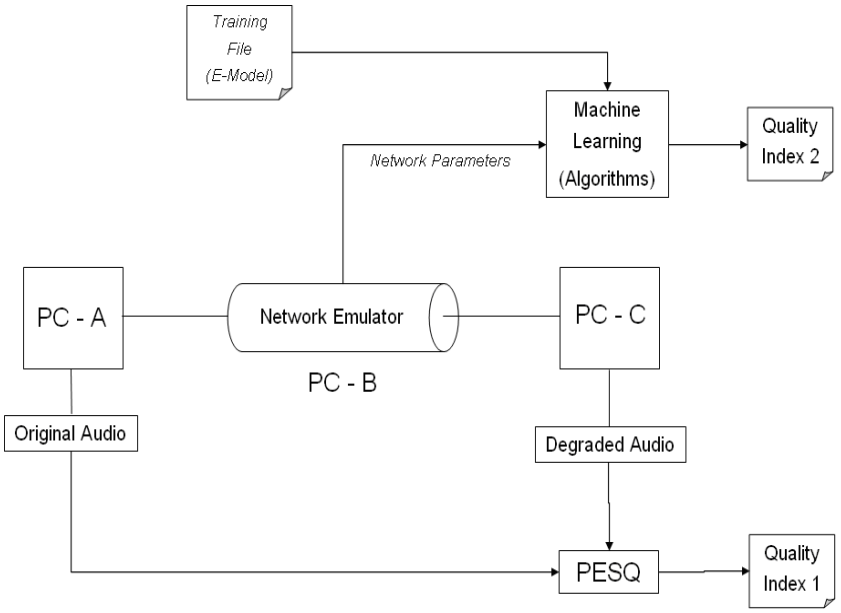

Figure 6: Test scenario.

In order to group ranges of R values and define QoS categories, we took the as reference model the classification presented by ITU-T Recommendation G.107. This classification model is presented in Figure 7, with the limit values for both R and MOS index.

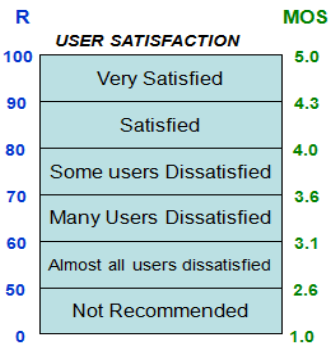

Figure 7: Levels of User Satisfaction of a VoIP communication.

The classification quality levels used herein only consider five categories; the last two categories, *Nearly All Users Dissatisfied* and *Not Recommended* in Figure 6, were grouped into one. These five categories are presented in Table X.

This categorization allows obtaining the file that will work as a training file to the algorithms that determines the quality of the communication. Table XI presents ten sample lines from the training file.

TABLE X: QoS levels used in the test scenario

Categories	R (Min. value)	R (Max. Value)
Very Satisfied (QoS_1)	90	94
Satisfied (QoS_2)	80	90
Some Users dissatisfied (QoS_3)	70	80
Many Users dissatisfied (QoS_4)	60	70
Nearly all Users Dissatisfied and Not Recommended (QoS_5)	0	60

TABLE XI: Codec and network parameters and the resulting QoS

Rate (kbps)	Delay (ms)	Bpl (%)	QoS
64	0	0	QoS_1
64	50	0	QoS_1
64	100	0	QoS_1
64	150	0	QoS_1
64	200	0	QoS_2
64	250	0	QoS_2
64	300	0	QoS_3
64	350	0	QoS_4
64	400	0	QoS_4
64	450	0	QoS_5

B. Determination of QoS using the Weka tool

In order to determine the QoS using the learning algorithms, Software Weka-version 3.7.4 [29] was used as a tool for data analysis method. This tool supports several algorithms, based on related works. The following four algorithms were used in the tests:

- Desicion Tree J48 - algorithm C4.5;
- Bayesian networks - Naive Bayes;
- Neural Networks - Multilayer Perceptron.
- Sequential Minimal Optimization - SMO.

With the training file, the cross-validation was analyzed and the values of the factor F (F-measure) for each algorithm tested are shown in Table XII.

TABLE XII: Values of F-measure for each algorithm and QoS

Algorithm	QoS-1	QoS-2	QoS-3	QoS-4	QoS-5
Trees J.48	0.99	0.96	0.96	0.98	0.97
Multilayer Perceptron	0.94	0.95	0.97	0.94	0.96
SMO	0.87	0.90	0.89	0.84	0.94
Bayes (Naives)	0.90	0.88	0.89	0.81	0.86

The values reached for the F factor (F-measure) were very high, whereas values greater than 0.7 are enough for network training. The decision tree algorithm obtained the best results, with higher F-measure value.

V. Experimental Results

In this work, 600 tests were conducted following the methodology explained. Each test considered a scenario configured with different parameters of network emulator. The larger the number of tests in learning techniques, the greater the validity of the results.

The results are presented in Figure 8, which depicts the highest value of success in determining the quality of service of VoIP communication for each learning algorithm used; the Decision Trees algorithm reaches 589 valid test results, that means, 589 results are concordant with the results obtained from PESQ. Also, the Multilayer Perceptron reached a high accuracy value with 570 satisfactory test results. SMO and Bayes algorithms reached 519 and 504 valid results, respectively. It is important to note that the performances of these four algorithms are concordant with the F-measure values presented in Table XII.

It worth noting that these results were obtained considering the range of MOS values of each QoS category and not a single value of index MOS.

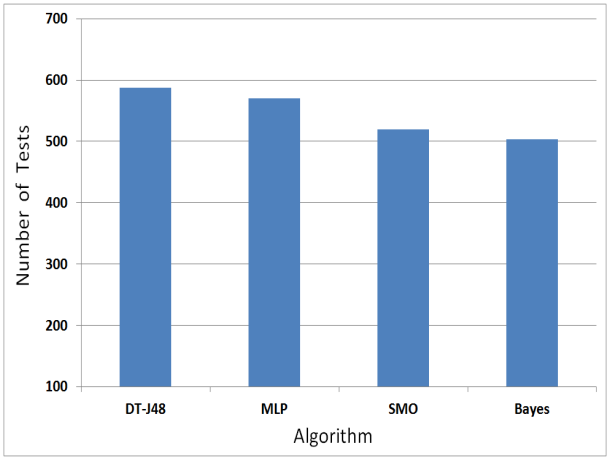

Figure 8: Algorithm results - Number of satisfactory test results according to PESQ algorithm.

VI. Conclusions and Future Work

The test results show that machine learning is a valid method to determine the quality of a VoIP communication in several network conditions and using different voice codecs, for this work, codecs G.711 and G.729. The good performance of the learning algorithms definitely depends on the initial file used to train these algorithms.

In this work, parameter values from a real IP network were used, and the test was performed in network with realistic parameters. Also, the confidence of the E-Model algorithm was tested to determine the values of the quality index for each scenario included in the training file. The best result was obtained by the Decision Tree algorithm that reached 98% of accuracy, which is a very high degree of reliability to determine the quality category of VoIP communication in relation to other works. The Multilayer Perceptron algorithm had a high level of success, too, reaching 95% accuracy. The Naive Bayes algorithm and SMO reached 86.5% and 84% of accuracy, respectively. The tests were performed in an emulated network with free softwares. For this reason, the implementation of the same scenario in other works is possible.

As future work, we intend to evaluate the quality of video services, for example, streaming video, creating a training file based on ITU-T Recommendation P.930 [30]. Also, different techniques of network resource allocation will be studied based on predictions of quality services of a future period of time, the goal of this idea being to improve the quality of a specific service.

Acknowledgments

The authors thanks University of São Paulo for the motivation to researches in the area of Computer and Telecommunication Systems. This work was supported by FAPESP (The State of São Paulo Research Foundation - Brazil). FAPESP project number: 2011/12724-8.

References

[1] D. Z. Rodríguez, L. R. Rosa, and G. Bressan, *Predicting the Quality Level of a VoIP Communication through Intelligent Learning Techniques*. The Seventh International Conference on Digital Society (ICDS), pp. 42-47, Nice, France, 2013.

[2] ITU-T Rec. P.800, *Methods for subjective determination of transmission quality*. Aug. 1996. Disponível em: www.itu.int/rec/T-REC-P.800/en, 08.12.2013.

[3] ITU-T Rec. P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*. http://www.itu.int/rec/TREC-P.862/en, 08.12.2013.

[4] ITU-T Rec. P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*, May. 2004. Disponível em: www.itu.int/rec/T-REC-P.563/en, 08.12.2013.

[5] ITU-T Rec. G.I07, *The E-model, a computational model for use in transmission planning*, International Telecommunications Union, Mar. 2005.

[6] L. A. R. Yamamoto and J. G. Beerends, *The impact of network performance parameters on the end-to-end perceived speech quality*, Expert ATM Traffic Symposium, Sep. 1997.

[7] S. R. Broom, *VoIP quality assessment: tracking account of the edge device*, IEEE Trans. on Audio, Speech and Language Processing, vol. 14, pp. 1977-1983, Nov. 2006.

[8] R. Eisinger. *Adaptação dinâmica de vídeo*. Master's thesis, Apr. 2007.

[9] A. Doulamis and G. Tziritas. *Content-based video adaptation in Low/Variable bandwidth communication networks using adaptable neural network structures*. In The 2006 IEEE International Joint Conference on Neural Network Proceedings, pages 4037-4044, Vancouver, BC, Canada, 2006.

[10] L. Sun and E. C. Ifeachor, *Perceived speech quality prediction for voice over IP based networks*, IEEE International Conference on Communications, vol. 4, pp. 2573-2577, Apr. 2002.

[11] S. Mohamed, F. Cervantes-Perez, and H. Afifi, *Integrating networks measurements and speech quality subjective scores for control purposes*, INFOCOM 2001 Proceedings, vol. 2, pp. 641-649, Apr. 2001.

[12] G. Corrigan, N. Massey, and O. Schnurr, *Transition-based speech synthesis using neural networks*, ICASSP 2000 Proceedings, vol. 2, pp. 945-948, Jun. 2000.

[13] M. M. Meky and T. N. Saadawi, *Prediction of speech quality using radial basis functions neural networks*, IEEE Proceeding on Computers and Communications, pp. 174-178, Jul. 1997.

[14] ITU-T Rec. G.711, *General Aspects of Digital Transmission Systems Terminal Equipments - Pulse Code Modulation (PCM) of Voice Frequencies*. 1972. http://www.itu.int/rec/T-REC-G.711/en, 08.12.2013.

[15] ITU-T Rec. G.729, *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*, Janeiro 2007. http://www.itu.int/rec/T-REC-G.729/en, 08.12.2013.

[16] S. Lingfen, *Perceived speech quality prediction for voice over IP-based networks*, IEEE International Conference on Communications, vol. 4, pp. 2573-2577, 2002.

[17] R. Jiuchun, D. Mao, and Z. Wang, *A neural network based model for VoIP speech quality prediction*, Proceedings of the 2nd International Conference on Interaction Sciences, pp. 1244-1248, 2009.

[18] J. R. Quinlan, *Induction of decision trees, Machine Learning*, 1986, vol. 1, issue 1, pp. 81 - 106.

[19] J. Stutz, *Bayesian Classification Theory*, Techinical Report, 1991.

[20] F. Rosenblatt, *Perceptron Simulation Experiments*, Proceedings of the IRE, vol. 48, pp. 301 - 309, 1960.

[21] S.-I. Horikawa, *On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm*, IEEE Transactions on Neural Networks, vol. 3, pp. 801 - 806, 1992.

[22] J. C. Platt, *Advances in kernel methods*. Cambridge, MA, USA: MIT Press, 1999.

[23] ITU-T Rec. P.861, *Objective quality measurement of telephone-band*. http://www.itu.int/rec/TREC-P.861/en, 08.12.2013.

[24] L. Ding, *Speech quality prediction in VoIP using the extended E-model*, IEEE Global Telecommunications Conference, pp. 3974 - 3978, vol.7, 2003.

[25] *Softphone Myphone*, http://myphone.sourceforge.net/, 08.12.2013.

[26] *Wireshark*, http://www.wireshark.org/download.html, 08.12.2013.

[27] *VRS - Recording System*, http://www.nch.com.au/vrs/index.html, 08.12.2013.

[28] *NETEM - Network Emulator*, http://www.linuxfoundation.org/collaborate/workgroups/networking /netem, 08.12.2013.

[29] Weka, *Weka 3: Data Mining Software in Java*. http://www.cs.waikato.ac.nz/ml/weka/, 08.12.2013.

[30] ITU-T Recommendation-P.930, *Principles of a reference impairment system for video*, Geneva, Aug. 1996.