Improving the performance of a non-intrusive metric of voice quality assessment considering IP network parameters

Carlos Henrique Pereira, Rodrigo Dantas Nunes, Renata Lopes Rosa, Thomaz Chaves de Andrade Oliveira and Demóstenes Zegarra Rodríguez

***Abstract*—In a voice-over-IP (VoIP) communication there are different factors that degrade the voice signal quality, and these degradations need to be evaluated in real time. In this scenario, the non-intrusive metrics such as the ITU-T P.563 recommendation are the most indicated. However, this metric has not a suitable performance if compared with intrusive metrics. In this study, an adjustment function is proposed to improve the P.563 metric's performance considering the packet loss parameter. The results show that the performance in relation to ITU-T P.862 was improved, obtaining a Pearson- correlation-coefficient and a maximum error of 0.99 and 0.40, respectively.**

Keywords—VoIP, Speech Quality, objective metrics, P.563, P.862, P.800, crowdsourcing.

I. INTRODUCTION

The speech communication quality is a determinant factor in user fidelity with telecommunication services operator. For this reason, the study of evaluation methods for speech quality is of great relevance. These methods can generate an index of speech quality, which can be used to improve a communication system's performance, increasing the user's quality of experience (QoE).

In general, these methods might be classified into two major groups: subjective and objective. Subjective methods are based on audition tests conducted in a controlled laboratory where volunteers follow the procedures that the test's supervisor establishes. An example of this subjective method is given by the ITU-T P.800 recommendation [1]. The subjective methods are the most exact and also serve to determine the performance of an objective metric [1].

There are different types of metrics or objective estimators of voice quality, each one focused on specific purposes or services. There are metrics called non-intrusive, because they just analyze the degraded signal voice at a specific point and / or in the end point of the connection [1]. On the other hand, there are other assessment metrics called intrusive metrics, such as ITU-T P.862 [2] and POLQA [3] recommendations, that provide better performance in relation to the results of subjective tests. However, they need a reference speech signal to compare with the degraded speech signal; thus, they might not be applicable in real-time services.

Therefore, considering a VoIP service, the P.563 [4] non-intrusive metric installed at the user's terminals could be of great interest to assess the speech communication quality. Although there are many studies which shows that the performance provided by this metric in VoIP service can be improved [5], [6] and [7]. This is due to the intrinsic characteristics of the algorithm that implements the P.563 metric, which is mainly based on the analysis of the vocal tract [4], and is not planned to consider external factors, such as, packet losses in an IP network.

Considering this limitation, Abareghi et al. [8] proposed the addition of a new distortion class to the P.563 algorithm in order to contemplate network conditions in the evaluation process. Other researches [9] and [10] propose new objective and non-intrusive methods, in which the first is based on statistic learning method, and the second, is based on Fuzzy Gaussian Mixture Model (FGMM) and Fuzzy Neural Network (FNN).

According to [11], the aspects of user's QoE are a vital factor to ensure the customer's satisfaction, and the packet loss in IP communication is a key point in this question. Thus, the study of a packet loss model is the main objective of researches, such as [11] and [12].

Similarly to the researches cited above, this study also needs to simulate the packet loss in an IP network in order to obtain the degraded signal at the end point of the communication. For this purpose, the Wav2Rtp [13] software was utilized, which is based on two-state Markov chain model [11], [12], [14]- [15], which is a widely used model for this type of simulation.

The main objective of this study is to propose a function based in the packet loss parameter, which is inserted into the P.563 algorithm, this is accomplished by adjusting the Mean Opinion Score (MOS) index. In order to get the mathematical model of this function, tests with different scenarios of packet loss and different speech signals were executed, obtaining for each scenario a MOS index from both ITU-T P.563 and 8.62 recommendations. Once the proposed model is determined, new test scenarios were performed and the results of the proposed model were compared with subjective assessment scores. Thus, the proposed model is a hybrid metric because considers both the voice signal and network parameters. In the validation tests of the proposed solution, a better performance of P.563 was observed in the VoIP service.

In this context, this work is divided in: Section II, in which the main methodologies of speech quality assessment are presented with emphasis in P.563 metric; Section III, that

Carlos Henrique Pereira, Rodrigo Dantas Nunes, Thomaz Chaves Andrade de Oliveira and Demóstenes Zegarra Rodríguez, Department of Engineering and Department of Science Computing, Federal University of Lavras, Lavras-MG, Brazil; and Renata Lopes Rosa, Departmet of Electrical Engineering, University of São Paulo, Brazil. E-mails: carlos@dcc.ufla.br, rdantasnunes@posgrad.ufla.br, rrosa@usp.br, thomaz.oliveira@dcc.ufla.br, demostenes.zegarra@dcc.ufla.br.
Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), process no.2012/24789-0. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) process no. 307633/2011-0.

describes the proposed model of the adjustment function; Section IV, where the test's methodology is presented; Section V will present the obtained results, and finally, Section VI are the conclusions.

II. Speech Quality Assessment Methods

The voice quality assessment tests can be executed following objective or subjective methods. The first one uses an algorithm, and the second one is based on user's opinion.

A. Subjective Methods

The ITU-T P.800 recommendation describes methods and procedures for conducting subjective assessment of quality of transmission. The subjective evaluations of equipment and telecommunications systems should at first be conducted using only listeners or conversational methods of subjective tests [6]. These methods include the following tests:

- Conversation opinion tests: Laboratory conversation tests are preferred, as far as possible, to reproduce, in the controlled laboratory, the actual service condition experienced by telephone customers. It is important that the simulated conditions in these tests are correctly specified, configured and accurately measured before and after each experiment [1].
- Listening opinion tests: It is not expected that these tests reach the same standards of realism that conversational tests and the restrictions are thus less severe in some aspects. The recommended test method for listening tests is the Absolute Category Rating (ACR), which is in conformance with the Category Judgment method recommended for conversation tests. The category scores are applied to a small group of unrelated sentences, and each one passed through a series of standard procedures. This is a method already solidified and has been applied to analog and digital telephone connections, and telecommunications devices, such as digital codecs [1].
- Interview and survey tests: If the rather large amount of effort needed is available, and the effort is justified, the transmission quality can be determined by "service observations". Recommended ways of performing this test, including questions to be asked to the interviewing customers, are given in P.82 recommendation. To maintain a high degree of precision is necessary to execute at least 100 interviews per condition. Although it has as disadvantage the fact of having little control over the detailed characteristics of the telephone connection, this method provides a global appreciation of how the equipment performs in a real setting [1].

Several principles are evaluated in this test and each gets a score of 1 to 5, where 1 represents the worst case and 5 represents the best. The scores are given by the participants are computed and the average of the values is calculated, which is called MOS. Table I presents the scale for the signal quality assessment, used in conversational tests and audition.

In recent years, a new approach of subjective tests for multimedia signal quality assessment is gaining popularity. This new approach takes advantage of the *crowdsourcing* concept, in which several workers perform a specific task and receive a monetary compensation. The main problem in conducting subjective tests in laboratories is the time consumed and the cost related to the infrastructure required. Currently, some researches performed quality assessment of multimedia applications via the Internet through remote assessors. Nowadays, there are some commercial solutions that have a user's database with users that are registered as workers. Also, this solution accepts different task descriptions to be performed by the workers. Moreover, some crowdsourcing solutions use social networks to perform some tasks without any payment. In this work, the crowdsourcing methodology is used to validate the performance of our proposed solution.

TABLE I. MOS Scale

Signal Quality	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

B. Objective Methods

An objective method determines an MOS index using an algorithm, for example, P.862 [2] and P.563 [4], and these metrics can be classified as intrusive and non-intrusive, respectively.

According to [4], a method is considered intrusive when a copy of the speech signal before the transmission is necessary in order that it can be compared to the degraded signal in the other transmission extremity. Conversely, a method is considered non-intrusive when only the degraded speech signal is needed. Fig. 1 demonstrates the difference between these two distinct approaches.

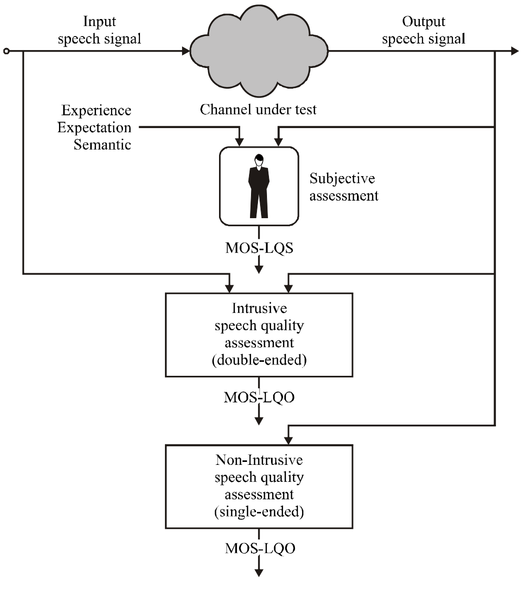

Fig. 1. Difference between intrusive and non-intrusive models [4].

The P.563 algorithm is applied in the voice quality prediction without a separate reference signal, and for this reason, is indicated for the real-time monitoring of communication networks and for the evaluation of unknown voice sources at the final extremity, or at a given point of a phone connection [4].

On the other hand, the P.862 algorithm, known as Perceptual Evaluation of Speech Quality (PESQ) compares an original signal X(t) with a degraded signal Y(t) that is the result of X(t) after passing through the communication system. The PESQ output is a prediction of the perceived quality that would be given by assessors in a subjective test [2].

III. Proposed Model

A solution that uses the proposed hybrid model is represented by Fig. 2. This model enables the calculation of a MOS index more correlated with the MOS values obtained from the P.862 intrusive method, which is validated with subjective tests.

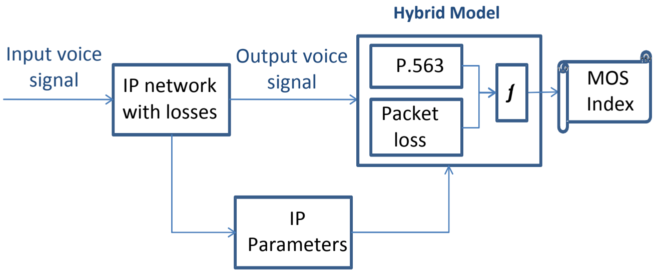

Fig. 2. Solution architecture proposal based on Rec. ITU T P.563 and the packet loss parameter.

The proposed hybrid model is based on the MOS index obtained from the P.563 algorithm, which analyzes a speech signal segment and the packet loss rate parameter. Thus, the obtained MOS index by P.563 is adjusted by a *f* function, whose output value approximates the results obtained by P.862.

The *f* function is modeled through experimental tests, in which twenty seven audio files extracted from a reference database were utilized. The reference database is named ANITA (Audio Enhancement Database In **S**ecured Telecom Applications) [16], and it considers different speakers, including both male and female genres.

Each audio file was degraded simulating network package losses. This was accomplished by utilizing the wav2rtp tool. The percentage of packet losses in each file can be controlled and thirty packet losses probabilities were generated in this process, which are: 0.5% up to 10% with steps of 0.5 and probabilities of 11% up to 20% with 1% steps. The wav2rtp tool was performed 100 times for each file considering a standard deviation of 2.5%. Consequently, a total of 810 degraded files were analyzed by P.563 and P.862 algorithms. With the obtained results the *f* adjustment function was achieved, for example, for the scenario of n% of packet loss, considering the audio file 1 (Arq-1) the following relation is obtained:

$$f_{Arq-1}^{n\%} = \frac{MOS(P.862_{Arq-1}^{n\%})}{MOS(P.563_{Arq-1}^{n\%})} \qquad (1)$$

In which, *MOS(P.862)* and *MOS(P.563)* represent the MOS index values obtained from the P.862 and P.563 algorithms, respectively. Thus, the value of $f_{mean}^{n\%}$ is the arithmetic average of the *f* values obtained for each of the 27 degraded files with n% packet loss. As stated before, 30 possible values for *n* were considered, *f'* discrete function is defined as $f' = [f_{mean}^{0.5\%}, f_{mean}^{1\%}, \dots, f_{mean}^{n\%}]$.

In order that *f '* is not restricted only for 30 cases of packet loss, *f ''* has been used to model the function *f (n)* using the linear regression approach, with the following polynomial function:

$$f(n) = \alpha \cdot n^3 + \beta \cdot n^2 + \gamma \cdot n + D \qquad (2)$$

IV. Tests Methodology

A. *Voice Data Base*

The voice signals used in the tests were taken from the database ANITA that contains voice files in different languages, with native and non-native speakers, and includes men and women recordings in normal, stress and panic conditions. These recordings were made in a laboratory environment with acoustic considerations [16]. In addition, ANITA has files recorded considering different noise source, such as wind, traffic, and siren. The recordings were stored in .wav files with 16 kHz Mono, a 16-bit rate, except for the siren .wav file that was recorded in stereo [16].

B. *Packet Loss Model*

The primary works about loss or error modeling occurred in the mid-1960s, and was based on errors related to telephone channels [17]. From then until the present date, several studies, such as [14], [18]-[20] contributed creating simulation models for packet loss, and other works use them for VoIP test scenarios [21]. Therefore, this work used the model described in [19] known as Gilbert-Elliott Model, which corresponds to the Markov chain of two states [14].

According to Yajnik et al. [14], the loss process in a Markov chain of two states is modeled as discrete-time. The current state, X_i of a stochastic process depends only of its previous value, that is, X_{i-1}. The author also states that this model is unlike to the Bernoulli model and it is able to capture the dependence between consecutive losses, but it has an additional parameter. The two parameters, p and q, are the probabilities of transition between the two states described in (3) and (4) as follows:

$$p = P[X_i = 1 \mid X_{i-1} = 0], \;\; \text{e} \;\; q = P[X_i = 0 \mid X_{i-1} = 1] \qquad (3)$$

The probability estimators of p and q, for a sample are:

$$\hat{p} = n_{01} / n_0 \,; \quad \hat{q} = n_{10} / n_1 \qquad (4)$$

In (4), for the observed time sequence, n_{01} is the number of times where 1 follows 0, and the parameter n_{10} is the number of times that 0 follows 1. Thus, n_0 is the number of 0s and n_1 is the number of 1s.

The distribution of successful execution for the application of this model is $f(j) = \hat{p}(1-\hat{p})^{j-1}$ for $j = 1,2,\dots\infty$ and the distribution for the loss of execution is $f(j) = \hat{q}(1-\hat{q})^{j-1}$ for $j = 1,2,\dots\infty$ [14].

Thus, the implemented algorithm in Wav2Rtp tool stores to begin the execution, the input parameters that are: the probability of loss of the current package when the previous packet was lost (a), and the probability of loss of the current packet when the previous packet was not lost (b). After it validates the informed values by the user for these probabilities, is not allowing values less than 0 or greater than 1. Each new package, the algorithm retrieves the current state of the file being degraded, then checks whether the last packet was lost, if so, uses the value of (a), and if not, the value of (b). In sequence, generates a random number, and if this is less than the value of the probability previously selected, causes the loss of the current package. This process is repeated until the entire file is processed.

C. *Tests Description*

As already described, the ANITA audio files were degraded by the Wav2Rtp software. Thus, from a voice file named original, degraded versions of the same voice with different packet loss rates could be obtained, and then, these degraded files were evaluated by the software of the ITU-T recommendations P.862 [2] and P.563 [4] to obtain the MOS indexes. It was necessary to resample the ANITA audio files from 16 kHz [16] to 8 kHz since Wav2Rtp only works with this sample rate. For this process the Audacity software version 2.0.6 [22] was used.

After this processing the Rec. P.862 and P.563 software were processed in these files, and the MOS values were obtained.

However, considering the huge mechanical work of repeating all these commands for each of the various ANITA database files, realized the need of develop a software whose purpose is to automate the execution of Wav2Rtp, P.862 and P.563. This software was developed using Java language with Swing API in aggregation with the Database Management System (DBMS) MySQL, and the tool for generating reports iReport.

In summary, this software receives the home directory of the ANITA database, a textual ID to be assigned to this execution, the speaker gender, the codec, a list of loss rates, and a value in milliseconds that determines the interval between executions of Wav2Rtp. Finally, general data relating to the used configuration and specific data from each file are stored on the software database.

At the end, a report is generated with different packet loss rates and the MOS values calculated for each file, allowing the analysis of results. Thus, new scenarios with different packet loss rates can be created quickly and easily.

With all this information available in the database, the analyses process even gained in dynamics, considering the fact that through the Structured Query Language (SQL) of DBMS was possible at any time calculations and analyses without the necessity of executing the software again to evaluate and index generation.

V. RESULTS

Considering that the current literature, that highlights that Rec. P.862 is one of the most reliable methods for assessing speech quality, the preliminary tests were performed comparing the obtained MOS values through P.563 and P.862 algorithms. As an example, Fig. 3 shows the results for the file "m_27_en_c_se06" from ANITA database. Other files in the same database were analyzed with different results, but with a similar pattern of behavior.

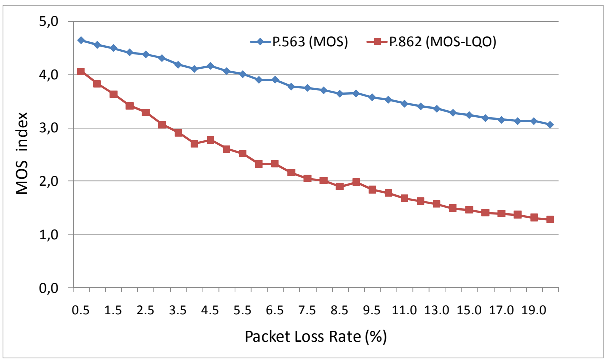

Fig. 3. MOS indexes obtained through P.563 and P.860 algorithms for 30 scenarios of packet loss

Using (1) the average values for each packet loss scenario were obtained: $f^{''}=$ [1.02; 0.98; 0.94; 0.91; 0.88; 0.84; 0.81; 0.79; 0.76; 0.74; 0.73; 0.71; 0.70; 0.68; 0.67; 0.65; 0.64; 0.63; 0.62; 0.61; 0.60; 0.58; 0.57; 0.56; 0.55; 0.54; 0.53; 0.53; 0.52; 0.52].

With the values of $f^{''}$, the function *f(n)* was modeled using (2), where α = -2 x 10^{-5}; $\beta = 0.001$; $\gamma = -0.043$ and $D = 1.059$.

With the *f(n)* adjustment function already defined, the initial values obtained by P.563 were reassessed, as shown below:

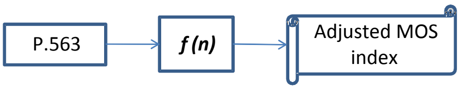

Fig. 4. The adjusted MOS index obtained by P.563 and the proposed *f(n)*

The results of *f(n)* performance evaluation regarding to P.862 results, considering the average MOS index of all evaluated audio files (30 impairment scenarios with packet loss rate) are presented in Table II.

TABLE II. PERFORMANCE EVALUATION OF THE PROPOSE FUNCTION *F(N)*

	Pearson Correlation Coefficient (PCC)	Maximum Absolute Error
P.563 original vs. P.862	0.9606	1.392
P.563 adjusted by *F(n)* vs. P.862	0.9973	0.406

Fig. 5 shows the discrete values of *f ’* and the proposed *f(n)*. The value of the coefficient of determination (R^2) between these two functions is 0.998, which is a highly reliable value.

It is worth noting that the function *f(n)* can accept different values of packet loss rate, which are monitored and extracted from a real IP network.

Also, the solution proposed can be used to evaluate real time services as VoIP, and also it has low complexity; therefore, consumes low processing resources considering the characteristic of current electronic devices.

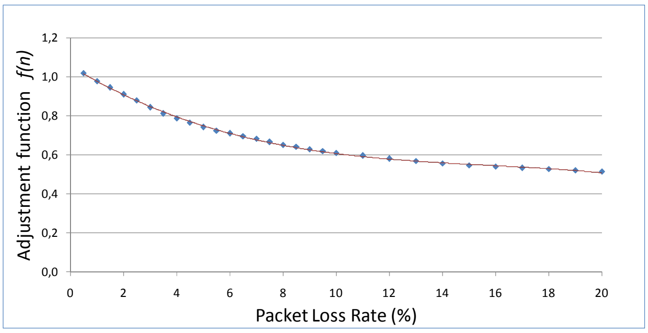

Fig. 5. Adjustment function based on the MOS indexes obtained through P.563 and P.860 algorithms for 30 scenarios of packet loss rate

In order to validate the performance of the proposed solution, three additional files of ANITA database were evaluated These files are different from the used to model *f(n)*. Each file was degraded with 3 different packet loss rate scenarios, which were of: 2%, 5% and 11%, obtaining nine different voice files to be evaluated. These remote subjective tests were performed using a *crowdsourcing* method, more specifically, using a commercial platform. For this, a web interface was built, in which the test instructions were specified and the voice files to be evaluated were uploaded.

Each voice file has been evaluated by 60 remote users, due to the fact that the crowdsourcing platform establishes 30 as the minimum number of participants for each campaign; thus, two campaigns were launched for each voice file. The results obtained are presented in Fig. 6.

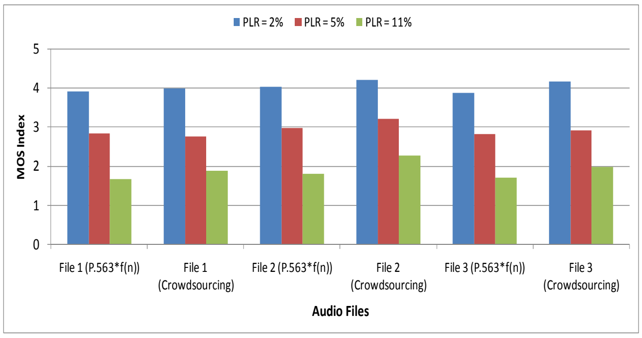

Fig. 6. Performance evaluation of the solution proposed regarding subjective tests using 9 impairment scenarios of audio files with PLR

As can be observed from Fig. 6, the MOS index granted by the remote users are similar to the values estimated by our proposed solution. In this case, the maximum error was of 0.47, considering the same 5-point MOS scale. It is worth noting that there are few samples to determine a reliable PCC value.

VI. Conclusions

In this work, a method to improve the P.563 algorithm's performance was proposed based on a function that adjusts the output value P.563. This was accomplished by simulating the packet loss effect in an IP network for VoIP communication. The results can be considered satisfactory, since the proposed method considering the P.862 results as reference reached an PCC and a maximum error of 0.99 and 0.40, respectively. Furthermore, remote subjective tests were conducted, and their results also demonstrated the high performance reached by the solution composed by the *f(n)* function and P.563 algorithm.

Also, the proposed methodology can be easily applied in real time services, and it has low cost in processing voice signals considering the current electronic devices.

Acknowledgements

The authors thank both the Signal Processing Laboratory and Network Architecture and Computer Laboratory of the Electrical Engineering Department from University of São Paulo, and also both Engineering and Computer Science Department from the Federal University of Lavras.

References

[1] ITU-T, "Methods for subjective determination of transmission quality," Tech. Rec. P.800, Geneva, Aug. 1996.

[2] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Tech. Rec. P.862, Geneva, Feb. 2001.

[3] ITU-T, "Perceptual objective listening quality assessment," Tech. Rec. P.863, Geneva, Nov. 2011.

[4] ITU-T, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Tech. Rec. P.563, Geneva, May. 2004.

[5] D. Picovici, A. Raja and C. Flanagan, "Real-Time, Non-intrusive Evaluation of VoIP," in Proc. EUROGP, 2007, pp. 217-228.

[6] T. H. Falk and Wai-Yip Chan, "Enhanced non-intrusive speech quality measurement using degradation models," in Proc. IEEE ICASSP, 2006, pp. I-I.

[7] W. Cherif, A. Ksentini, D. Negru and M. Sidibe, "A_PSQA: PESQ-like non-intrusive tool for QoE prediction in VoIP services," in Proc. IEEE ICC, 2012, pp.2124–2128.

[8] M. Abareghi, M. M. Homayounpour, M. Dehghan and A. Davoodi, "Improved ITU-T P.563 Non-Intrusive Speech Quality Assessment Method For Covering VOIP Conditions," in Proc. IEEE ICACT, 2008, pp. 354-357.

[9] D. S. Kim and A. Tarraf, "Enhanced perceptual model for non-intrusive speech quality assessment," in Proc. IEEE ICASSP, 2006, pp. I-I.

[10] J. Wang, Y. Zhang, Y. Song, S. Zhao and J. Kuang, "An improved non-intrusive objective speech quality evaluation based on FGMM and FNN," in Proc. IEEE CISP, 2010, pp. 3495-3499.

[11] O. Hohlfeld, R. Geib and G. Haßlinger, "Packet loss in real-time services: Markovian models generating QoE impairments," in Proc. IEEE IWQoS, 2008, pp. 261-270.

[12] Z. Li, J. Chakareski, X. Niu, G. Xiao, Y. Zhang and W. Gu, "Modeling and analysis of distortion caused by Markov-model burst packet losses in video transmission," IEEE Trans. Circuits Systems for Video Technology, vol. 19, pp. 917-931, Jul. 2009.

[13] Wav2Rtp Software. (2015, Apr 03). [Online]. Available: http://wav2rtp.sourceforge.net

[14] M. Yajnik, S. B. Moon, J. Kurose and D. Towsley, "Measurement and modeling of the temporal dependence in packet loss," in Proc. IEEE INFOCOM, 1999, pp. 345–352.

[15] X. Yu, J. W. Modestino and X. Tian, "The accuracy of Gilbert models in predicting packet-loss statistics for a single-multiplexer network model," in Proc. IEEE INFOCOM, 2005, pp. 2602-2612.

[16] Information Society Technologies, "Anita Reference Database Description," vol.2, EADS Telecom, 2003, pp. 06-27.

[17] Voip Troubleshooter. (2015, Apr 02). Packet Loss Burtiness. [Online]. Available: http://www.voiptroubleshooter.com/indepth/burstloss.html

[18] E. N. Gilbert, "Capacity of burst-noise channel," Bell System Technical Journal, vol. 39, pp. 1253–1265, Sep. 1960.

[19] E. O. Elliott, "A Model of the Switched Telephone Network for Data Communications," Bell System Technical Journal, vol. 44, pp. 89-109, Jan 1965.

[20] A. H. Blank and J. P. Trafton, "A Markov Error Channel Model", in Proc. Nat Telecomm Conference, 1973.

[21] D. Rodríguez, M. Arjona, "VoIP Quality Improvement with a Rate-determination Algorithm", in Proc. IWT, pp. 30-34, Brazil, Feb. 2009.

[22] Audacity. (2015, Apr 03). [Online]. Available: http://audacity.sourceforge.net