A Comprehensive Investigation on Image Caption Generation using Deep Neural Networks

ANDREZA PATRÍCIA BATISTA[1] LUCAS HILÁRIO DA COSTA[2]
DEMÓSTENES ZEGARRA RODRÍGUEZ[3]

Federal University of Lavras
Departamento de Engenharia - Engenharia de Sistemas e Automação, UFLA
[1]andreza.batista@estudante.ufla.br
[2]costa.lh@estudante.ufla.br
[3]demostenes.zegarra@dcc.ufla.br

Abstract. Currently, Voice over IP (VoIP) is one of the most used communication services, however, its quality is related to several external factors that cause various types of degradation of the voice signal, directly affecting the quality of experience (QoE) of users. In order to classify the quality of the voice signal transmitted in a VoIP communication affected by packet loss, two deep learning network models (DL - *Deep Learning*) were implemented. The models were developed using a deep neural network model (DNN), through which the analysis of the voice signal affected by the packet loss rate (PLR) of the degraded signals, so it was possible to classify them into four different classs according to the user's experience. Thus, two databases were prepared, each containing four distinct classs. One of these was prepared with the ITU-T P.862 recommendation database files with different packet loss rates, and the other database was prepared with the ITU-T P.501 recommendation files according to the index MOS of Mean Opinion Score (MOS) of each degraded file. The results obtained from the model for the database prepared by the packet loss rate was 94% accuracy in model validation, while the model results for the database prepared by MOS the result obtained was 91% of accuracy. In a comparison with the results obtained by the P.563 algorithm and the results obtained by the P.862 algorithm, it was possible to obtain an average of 53.21% accuracy for the P.563 algorithm in comparison with the classification results of the algorithm P.862. Through the results obtained, it can be concluded that the generated models were able to classify the packet loss rate and the MOS index in a non-intrusive way and with a great accuracy rate. Concluding that the generated models are able to determine the MOS of the degraded voice files more efficiently than the P.563 algorithm.

Keywords: VoIP, Voice Quality, ITU-T P.862, ITU-T P.563, ITU-T P.501, Deep Learning, Machine Learning

(Received November 21st,2022 / Accepted December 11st, 2022)

1 Introduction

According to the *Cisco Visual Networking Index Forecast* it is predicted that by 2022 the percentage of the world's population using the Internet will reach 60%, and consequently increase the number of devices connected to the network, from 2.4 to 3 ,6 devices per person . With the evolution of the existing network infrastructures, there has been a considerable improvement in the QoS available to the end user and this will increase even more with the development of new technologies, such as, for example, the fifth generation of communications networks (5G). As a result, the demand for multimedia services and user expectations are also increasing. This has led to a greater demand for more bandwidth resulting in more challenges for network operators with already limited resources.

According to [26], VoIP applications are becoming popular these days generating a lot of Internet traffic. Normally, VoIP traffic is carried over *User Datagram Protocol* (UDP), except when firewalls block UDP, in which case voice signal and signaling traffic is carried over *Transmission Control Protocol* (TCP).

Delays, packet loss and jitter (delay variation) directly affect the quality of the transmitted signal. These packet losses may be linked to the discarding of packets by congested routers and/or problems in the physical means of transport.

In order to evaluate the quality of the voice signal caused by network degradation factors [17, 18, 20, 11, 28, 27, 8, 16, 5, 10, 21], there are well-defined methods [15, 2, 1], which are specified through the technical recommendations of the *International Telecommunication Union - Telecommunication Standardization Sector* (ITU-T) which provides recommendations for several areas and purposes within the telecommunications context. These recommendations define how the voice quality will be evaluated, scoring the quality of communications [22, 23, 3, 19, 14], in addition, these recommendations are of great importance for monitoring the QoE of service users.

Voice quality assessment methods can be classified into subjective or objective methods. In the subjective method, a number of people are invited to evaluate the quality of the voice samples, this evaluation results in the MOS [13] index, which is the grade given to the voice quality by the user's experience. However, it is a time consuming and costly method to be used to continuously monitor system QoE in real voice communication scenario. Objective methods, on the other hand, use algorithms and attempt to approximately predict the user's QoE score, which would be given in subjective tests by individuals. In addition, objective methods are subdivided into two methods, intrusive and non-intrusive. The non-intrusive one needs only the degraded signal to determine the voice quality, while the intrusive one needs the original signal as a reference to evaluate the quality of the degraded signal.

According to the [12] recommendation, the objective and non-intrusive method is the most recommended for evaluating real-time communications such as VoIP. Seeking to analyze and improve these problems, an implementation of a DL [24] algorithm capable of evaluating and identifying voice quality in VoIP communications was developed in this work.

A typical *Deep Learning* algorithm usually includes two processes: an unsupervised learning step and a supervised parameter adjustment step [29]. Decision making in the neural network is done layer by layer, in which the following layers are able to make increasingly specific decisions in a learning process supervised by the chosen optimization algorithm. After training the neural network, the parameters are used to initialize an optimized network with respect to a supervised training criterion [7].

Recurrent Neural Networks, as well as the well-known multilayer networks MLP (*Multi-layer Perceptron*), are deep learning models par excellence and operate on sequences of vectors with the aim of approximating some function. Unlike neural networks of the *feedforward* type, recurrent networks allow the use of information from their own output delayed in time to help obtain better results, since their recurrent layers have feedback loops that allow keeping the information in memory over time. Recurrent Neural Networks (RNN) have been successfully used in language models [9], handwriting recognition and generation, speech recognition, among other applications.

On the other hand, common recurrent neural networks suffer from a serious problem: the vanishing or exploding gradient or *vanishing gradient problem*. It becomes quite difficult for a common recurrent neural network to solve problems that require learning long-term temporal dependencies, that is, involving many time delays. The occurrence of this is due to the gradient of the weight adjustment function that decays or increases exponentially with time, making it difficult to update the weights during the *backpropagation* phase.

The convolutional neural network is a class of deep neural network that exploits the strong local and spatial correlation in natural images, achieving great performance in the area of visual analysis. Recently, CNNs have been employed in the area of ââacoustic processing and have proven to be able to learn the spectro-temporal pattern of sound and differentiate it for [25] classification purposes. In recent years, several acoustic scene classification algorithms have been proposed where the most used algorithms include Support-Vector Machine (SVM), Gaussian Mixture Models (GMM), Hidden Markov Models (HMM - Hidden Markov Models), or the hierarchy of these methods.

In the first model, the database files were separated into 4 classes with different PLR rates. In the second model, the database files underwent an evaluation of the MOS index through the ITU-T P.862 recommendation algorithm and soon after the degraded audio files were separated into 4 classes according to the R-Factor Table of E-Model MOS index [6] and [4].

The main objective of this work was the development of a non-intrusive model for identifying and classifying voice quality in VoIP communication using Deep Learning. Through the specification and training of neural network algorithms, two models capable of classifying the degree of voice degradation through the MOS index and through the analysis of the packet loss rate of the voice signal were implemented.

2 Methodology

To carry out the work, different steps were taken to reach the final result, we can see in Figure 1 the steps taken in the project:

2.1 Language and Libraries

The neural network models were fully developed in the Python 3.6 language using the Ubuntu terminal with the help of Keras and Tensorflow APIs.

- Keras: It is a high-level API for neural networks, written in Python and uses TensorFlow, CNTK or Theano as a backend. It was designed with a focus on rapid development, being able to move from idea to result with as little effort as possible.

- TensorFlow: It is an open source software library for numerical computation using data flow graphs. The nodes in the graph represent mathematical operations, while the edges of the graph represent the multidimensional data arrays (tensors) communicated between them. The flexible architecture lets you deploy compute for one or more CPUs or GPUs on a desktop, server, or mobile device with a single API.

- LibROSA: LibROSA is a python package for music and audio analysis. It provides the necessary building blocks to create musical information retrieval systems.

- Python: Python is a high-level, interpreted, scripting, imperative, object-oriented, functional, dynamically typed, strong programming language. It was launched by Guido van Rossum in 1991.

- Dropbox: Dropbox is a service for storing and sharing files. It is based on the concept of "cloud computing", and was used to store the project's databases and algorithms.

- MATLAB R2017 (MATrix LABoratory): This is a high-performance interactive software focused on numerical calculation. MATLAB integrates numerical analysis, matrix calculation, signal processing, and graphing into an easy-to-use environment where problems and solutions are expressed just as they are written mathematically.

2.2 Dataset

For the creation of the databases, the databases of the ITU-T P.501 recommendation and the ITU-T P.862 recommendation were first downloaded, to prepare the databases a processing in MATLAB was carried out which simulates a given packet loss rate in the original voice file in the database.

2.2.1 PLR classs

For the packet loss classification model, after file processing, files were separated into four distinct classs according to their PLR rate. Where the classs were separated as follows:

- Class 1: PLR(0.5% + 1.0% + 1.5% + 2.0%) with a total of 1840 degraded files;

- Class 2: PLR(2.5% + 3.5% + 4.5% + 5.5%) with a total of 1840 degraded files;

- Class 3: PLR(7.0% + 8.0% + 9.0% + 10.0%) with a total of 1840 degraded files;

- Class 4: PLR(12.0% + 15.0% + 18.0% + 21.0%) with a total of 1840 degraded files;

To generate each class, the following procedure was performed: for each of the 20 audio files of the original base, processing was carried out with the 16 packet loss rates resulting in a total of 320 files for each execution of the algorithm; to increase the number of files and the diversification between them, the algorithm was executed 23 times, but this number does not represent any parameter. The result generated a total of 7360 degraded audio files, and as each class has the same number of degraded files and is formed by 4 different PLR rates, there are 1840 per class.

2.2.2 MOS classs

For the MOS index classification model, after processing the files, a new algorithm (also developed in MATLAB) was executed, capable of automating the execution of the P.862 algorithm for all degraded audio files, comparing them with their respective original audio files. The result was the MOS index referring to the packet loss found in the degraded audio file.

After acquiring these data, the files were separated into 5 class according to their MOS index, using the E-Model R-Factor table shown in Figure 2 as a reference.

However, it was detected that files with MOS index were not generated for class 5 where the MOS index is between 4.3 and 4.5, so it was adopted that files with indexes greater than 4.0 would be only one class, thus resulting in 4 classs for the indices found according to Figure 4.6.

- Class 1: MOS(1.00 to 3.09) with a total of 2500 degraded files;
- Class 2: MOS(3.10 to 3.59) with a total of 2500 degraded files;
- Class 3: MOS(3.60 to 3.99) with a total of 2500 degraded files;
- Class 4: MOS(4.00 to 4.50) with a total of 2500 degraded files;

2.3 Network Topology

The Deep Learning architecture used in the training and validation of the project was defined according to the number of characteristics of the files in the neural network input, for which a neural network model was defined as shown in Figure 3:

The generated model has the following characteristics:

- It has an input layer (I) with 96 neurons, where 96 is the number of features that will be used in the network, defined automatically by the algorithm.
- It has seven hidden layers (H) with 96 neurons each, recurrently connected according to the RNN's.
- It has an output layer (O) with 4 neurons, where each one represents an output of a class.

3 Results

To carry out the training and validation stages, a computer with a 2.3 Ghz I5 processor, 8GB of RAM and a GTX 1050 TI graphics card with 4GB of video memory was used, with an average algorithm execution time of 9 hours.

To carry out the tests, 80% of the audio files were used for training and 20% for model validation. To perform the extraction of features from the audio file, CNN's mainly generate spectrograms and analyze them as a pixel-by-pixel image.

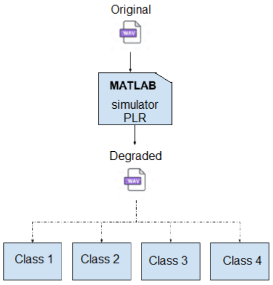
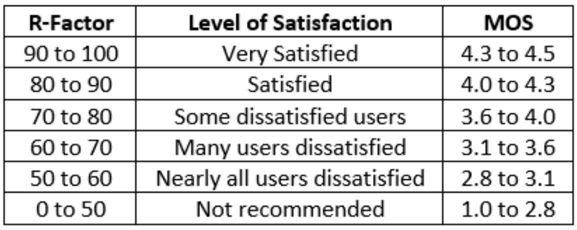


Figure 1: Methodology Stepes



R-Factor	Level of Satisfaction	MOS
90 to 100	Very Satisfied	4.3 to 4.5
80 to 90	Satisfied	4.0 to 4.3
70 to 80	Some dissatisfied users	3.6 to 4.0
60 to 70	Many users dissatisfied	3.1 to 3.6
50 to 60	Nearly all users dissatisfied	2.8 to 3.1
0 to 50	Not recommended	1.0 to 2.8

Figure 2: R-Factor table compared to the MOS index.

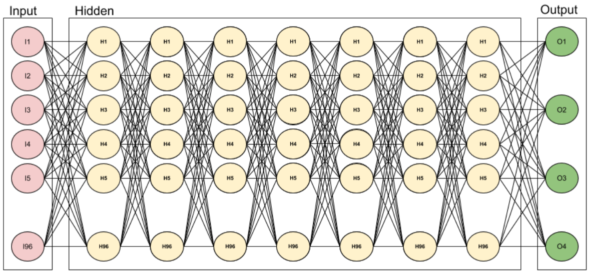


Figure 3: Deep Learning Model.

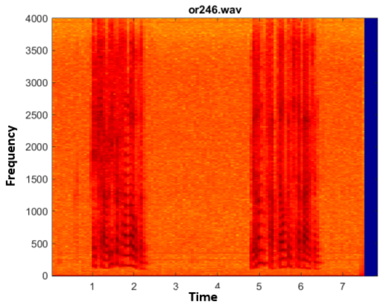


Figure 4: Spectrogram of the original audio file.

Figure 4 represents a spectrogram of an original file, where we can verify that the voice signal did not suffer any kind of degradation, in this the degradation is represented by the blue color. However, the blue color does not only represent the loss of the voice signal, it also represents the silence in the conversation.

However, the algorithm of the ITU-T P.862 recommendation has the ability to differentiate silences and losses in the voice signal, on the other hand, the algorithm of the ITU-T P.563 recommendation does not have this capability.

Figure 5 represents the spectrograms generated for the same audio file represented by Figure 4, but as we can see the degradation occurred by the defined packet loss rates represented a degradation of the voice signal for the different rates, and as the percentage packet loss increases, we can see that there is an increase in silences in the voice signal. The letters ”a” to ”p” respectively represent the values of PLR 0.5%, 1.0%, 1.5%, 2.0%, 2.5%, 3.5%, 4.5%, 5.5

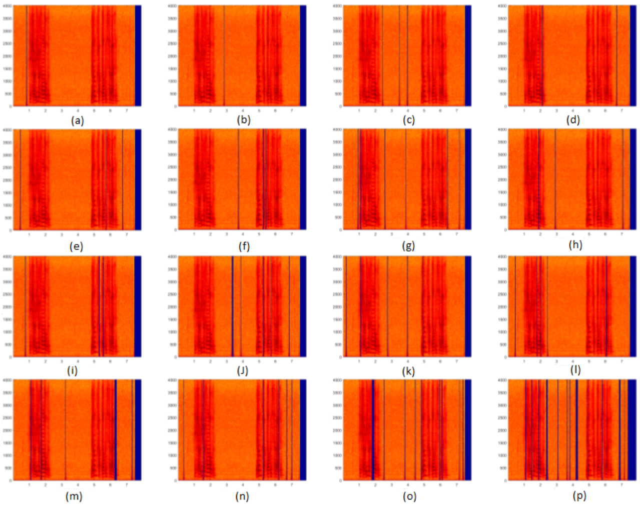


Figure 5: Spectrograms of degraded audio files for PLR values.

3.1 Results using the PLR model

The results obtained in the tests with the database prepared with the values ââof PLR, were obtained through the validation of the model using 20% of the total files of the database, that is, 1472 files, being 368 files for each defined class. The results are shown in confusion matrices generated for epoch numbers 10, 50, 100, 500 and 1000, and are presented in Tables I, II, III, IV and V.

Table I presents the data of successes and errors of the tests carried out for 10 training epochs, through it we can see that in the validation of the model trained with only 10 epochs we had a total of 230 class classification errors and a total of 1242 hits in the validation, obtaining an accuracy of 84.375% in model validation.

Table II presents the data of successes and errors of the tests carried out for 50 training epochs, through it we can see that in the validation of the model trained with 50 epochs we had a total of 175 class classification errors and a total of 1297 hits in the validation , obtaining an accuracy of 88.111%.

Table III presents the data of successes and errors of the tests performed for 100 training epochs, through it we can see that in the validation of the model trained with 100 epochs we had a total of 157 errors in class classification and a total of 1315 correct answers in the validation , obtaining an accuracy of 89.334%.

Table IV presents the data of successes and errors of the tests carried out for 500 training epochs, through it we can see that in the validation of the model trained with 500 epochs we had a total of 149 class classifica-

Table 1: Confusion matrix result for 10 epochs for PLR values

	Class 1	Class 2	Class 3	Class 4
Class 1	351	17	-	-
Class 2	11	317	40	-
Class 3	-	50	278	40
Class 4	-	-	72	296

Table 2: Confusion matrix result for 50 epochs for PLR values

	Class 1	Class 2	Class 3	Class 4
Class 1	353	15	-	-
Class 2	5	328	35	-
Class 3	-	30	306	32
Class 4	-	-	58	310

Table 3: Confusion matrix result for 100 epochs for PLR values

	Class 1	Class 2	Class 3	Class 4
Class 1	356	12	-	-
Class 2	10	319	39	-
Class 3	-	20	321	27
Class 4	-	-	49	319

Table 4: Confusion matrix result for 500 epochs for PLR values

	Class 1	class 2	class 3	class 4
Class 1	359	9	-	-
class 2	11	325	32	-
class 3	-	21	319	28
class 4	-	-	48	320

tion errors and a total of 1323 hits in the validation , obtaining an accuracy of 89.877%.

Table V presents the data of successes and errors of the tests carried out for 1000 training epochs, through it we can see that in the validation of the model trained with 1000 epochs we had a total of 89 class classification errors and a total of 1383 hits in the validation , obtaining an accuracy of 93.954%.

3.2 Results using the MOS model

The results obtained in the tests with the database prepared with the values ââof the MOS index were obtained through the validation of the model using 30% of the total files of the database, where each class contains a total of 750 files for validation, resulting in a total of 3000 files for model validation.

The results are shown in confusion matrices generated for the following epoch numbers 10, 50, 100, 500 and 1000, and are presented in Tables VI, VII, VII,IX and X.

Table VI presents the data of successes and errors of the tests carried out for 10 training epochs, through it we can see that in the validation of the model trained with 10 epochs we had a total of 341 class classification errors and a total of 2659 hits in the validation , obtaining an accuracy of 88.633%

Table VII presents the data of successes and errors of the tests carried out for 50 training epochs, through it we can see that in the validation of the model trained with 50 epochs we had a total of 272 class classification errors and a total of 2728 classification correctness , obtaining an accuracy of 90.933%.

Table VIII presents the data of successes and errors of the tests performed for 100 training epochs, through it we can see that in the validation of the trained model with 100 epochs we had a total of 262 class classification errors and a total of 2736 classification correctness , obtaining an accuracy of 91,200%.

Table IX presents the data of successes and errors of the tests carried out for 500 training epochs, through it we can see that in the validation of the model trained with 500 epochs we had a total of 182 class classification errors and a total of 2818 classification correctness , obtaining an accuracy of 93.933%.

Table X presents the data of successes and errors of the tests performed for 1000 training epochs, through it we can see that in the validation of the model trained with 1000 epochs we had a total of 106 class classification errors and a total of 2894 classification successes , obtaining an accuracy of 96.467%.

Table 5: Confusion matrix result for 1000 epochs for PLR values

	class 1	class 2	class 3	class 4
class 1	357	7	-	-
class 2	6	332	17	-
class 3	-	14	321	22
class 4	-	-	23	329

Table 6: Resultado da matriz de confusÃ£o para 10 Ã©pocas para valores de MOS

	class 1	class 2	class 3	class 4
class 1	674	76	-	-
class 2	58	657	35	-
class 3	-	48	669	33
class 4	-	-	51	659

Table 7: Confusion matrix result for 50 epochs for MOS values

	class 1	class 2	class 3	class 4
class 1	695	55	-	-
class 2	40	678	32	-
class 3	-	39	683	28
class 4	-	-	78	672

Table 8: Confusion matrix result for 100 epochs for MOS values

	class 1	class 2	class 3	class 4
class 1	699	51	-	-
class 2	36	680	34	-
class 3	-	37	681	32
class 4	-	-	72	678

Table 9: Confusion matrix result for 500 epochs for MOS values

	class 1	class 2	class 3	class 4
class 1	716	34	-	-
class 2	31	694	25	-
class 3	-	28	705	17
class 4	-	-	47	703

Table 10: Confusion matrix result for 1000 epochs for MOS values

	class 1	class 2	class 3	class 4
class 1	725	25	-	-
class 2	15	722	13	-
class 3	-	19	718	13
class 4	-	-	21	729

3.3 Comparison of P.563 vs P.862 algorithms

When performing the classification of the R-Factor MOS indexes with the results of the algorithms of the ITU-T P.563 and ITU-T P.862 recommendations, we obtained, according to Table XII, an average of 61.83% of hit when we compare the results obtained with the P.563 algorithm with the results obtained with the P.862 algorithm. The P.563 algorithm correctly classified the classes of 6183 files and failed the classes of 3817 files in the database.

By comparing the results obtained from the P.563 algorithm with the results obtained by the model, considering that the model was trained and validated according to the results of the MOS obtained by the ITU-T P.862 recommendation algorithm, we can verify that the model obtained a better performance than the P.563 algorithm, where an accuracy of 96.467% was obtained for the model and 61.83% for the P.563 algorithm.

4 Conclusion

Through the results achieved, it can be concluded that the proposed model meets the main objective of the project, which is to classify the packet loss rate and identify the MOS index through a Deep Learning algorithm, in which its hit rate was on average 94% for the model trained by the packet loss rate and 96.467% for the model trained by the MOS index of the degraded audio files. We can also highlight that through the results of the proposed model with the results of the algorithm of the ITU-T P.563 recommendation, it is noticed that the proposed model obtained a greater efficiency than the algorithm of the ITU-T P.563 recommendation, becoming a voice quality analysis model in a non-intrusive way with greater efficiency, being able to monitor VoIP transmissions, recognizing the packet loss rate in real time and running some correction tool so that the user experience is not affected.

References

[1] Affonso, E. T., Nunes, R. D., Rosa, R. L., Pivaro, G. F., and Rodríguez, D. Z. Speech quality assessment in wireless voip communication using deep belief network. *IEEE Access*, 6:77022–77032, 2018.

[2] Affonso, E. T., Rodríguez, D. Z., Rosa, R. L., Andrade, T., and Bressan, G. Voice quality assessment in mobile devices considering different fading models. In *2016 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 21–22. IEEE, 2016.

[3] Affonso, E. T., Rosa, R. L., and Rodríguez, D. Z. Speech quality assessment over lossy transmission channels using deep belief networks. *IEEE Signal Processing Letters*, 25(1):70–74, 2017.

[4] Bergstra, J. A. and Middelburg, C. Itu-t recommendation g. 107: The e-model, a computational model for use in transmission planning. 2003.

[5] de Almeida, F. L., Rosa, R. L., and Rodríguez, D. Z. Voice quality assessment in communication services using deep learning. In *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pages 1–6. IEEE, 2018.

[6] G.107, I.-T. R. The e-model: a computational model for use in transmission planning. June 2015. Acessado: 22 Abril 2019.

[7] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

[8] Jordane da Silva, M., Carrillo Melgarejo, D., Lopes Rosa, R., and Zegarra Rodríguez, D. Speech quality classifier model based on dbn that considers atmospheric phenomena. *Journal of Communications Software and Systems*, 16(1):75–84, 2020.

[9] Karpathy, A., Johnson, J., and Li, F. Visualizing and understanding recurrent networks. *CoRR*, 1506.02078, 2015.

[10] Nunes, R. D., Pereira, C. H., Rosa, R. L., and Rodríguez, D. Z. Real-time evaluation of speech quality in mobile communication services. In *2016 IEEE International Conference on Consumer Electronics (ICCE)*, pages 389–390. IEEE, 2016.

[11] Nunes, R. D., Rosa, R. L., and Rodríguez, D. Z. Performance improvement of a non-intrusive voice quality metric in lossy networks. *IET Communications*, 13(20):3401–3408, 2019.

[12] P.563, I.-T. R. Single-ended method for objective speech quality assessment in narrow-band telephone applications. Apr. 2004.

[13] P.800, I.-T. R. Methods for subjective determination of transmission quality. Aug. 1996.

[14] Rodríguez, D. Z., Abrahao, J., Begazo, D. C., Rosa, R. L., and Bressan, G. Quality metric to assess video streaming service over tcp considering temporal location of pauses. *IEEE Transactions on Consumer Electronics*, 58(3):985–992, 2012.

[15] Rodríguez, D. Z. and Bressan, G. Video quality assessments on digital tv and video streaming services using objective metrics. *IEEE Latin America Transactions*, 10(1):1184–1189, 2012.

[16] Rodríguez, D. Z., Carrillo, D., Ramírez, M. A., Nardelli, P. H. J., and Möller, S. Incorporating wireless communication parameters into the e-model algorithm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:956–968, 2021.

[17] Rodríguez, D. Z., da Silva, M. J., Silva, F. J. M., and Junior, L. C. B. Assessment of transmitted speech signal degradations in rician and rayleigh channel models. *INFOCOMP Journal of Computer Science*, 17(2):23–31, 2018.

[18] Rodríguez, D. Z. and Junior, L. C. B. Determining a non-intrusive voice quality model using machine learning and signal analysis in time. *INFOCOMP Journal of Computer Science*, 18(2), 2019.

[19] Rodríguez, D. Z., Rosa, R. L., Alfaia, E. C., Abrahão, J. I., and Bressan, G. Video quality metric for streaming service using dash standard. *IEEE Transactions on broadcasting*, 62(3):628–639, 2016.

[20] Rodríguez, D. Z., Rosa, R. L., Almeida, F. L., Mittag, G., and Möller, S. Speech quality assessment in wireless communications with mimo systems using a parametric model. *IEEE Access*, 7:35719–35730, 2019.

[21] Rodríguez, D. Z., Rosa, R. L., and Bressan, G. A billing system model for voice call service in cellular networks based on voice quality. In *2013 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 89–90. IEEE, 2013.

[22] Rodríguez, D. Z., Rosa, R. L., Costa, E. A., Abrahão, J., and Bressan, G. Video quality assessment in video streaming services considering user preference for video content. *IEEE Transactions on Consumer Electronics*, 60(3):436–444, 2014.

[23] Rodríguez, D. Z., Wang, Z., Rosa, R. L., and Bressan, G. The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over http. *EURASIP Journal on Wireless Communications and Networking*, 2014(1):1–15, 2014.

[24] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[25] Shu, H., Song, Y., and Zhou, H. Time-frequency performance study on urban sound classification with convolutional neural network. In *TENCON 2018-2018 IEEE Region 10 Conference*, pages 1713–1717. IEEE, 2018.

[26] Sinam, T., Singh, I. T., Lamabam, P., Devi, N. N., and Nandi, S. A technique for classification of voip flows in udp media streams using voip signalling traffic. In *2014 IEEE International Advance Computing Conference (IACC)*, pages 354–359, Feb 2014.

[27] Terra Vieira, S., Lopes Rosa, R., Zegarra Rodríguez, D., Arjona Ramírez, M., Saadi, M., and Wuttisittikulkij, L. Q-meter: Quality monitoring system for telecommunication services based on sentiment analysis using deep learning. *Sensors*, 21(5):1880, 2021.

[28] Vieira, S. T., Rosa, R. L., and Rodríguez, D. Z. A speech quality classifier based on tree-cnn algorithm that considers network degradations. *Journal of Communications Software and Systems*, 16(2):180–187, 2020.

[29] Yan, W., Tang, D., and Lin, Y. A data-driven soft sensor modeling method based on deep learning and its application. *IEEE Transactions on Industrial Electronics*, 64(5):4237–4245, May 2017.

Table 11: Result of the confusion matrix of the comparison of the results of the MOS index of the executions of the ITU-T P.563 and ITU-T P.862 algorithms.

	C1-P.862	C2-P.862	C3-P.862	C4-P.862
C1-P.563	4609	2381	350	104
C2-P.563	11	1559	35	4
C3-P.563	0	48	899	0
C4-P.563	0	0	0	0