Evaluation of objective audio quality metrics in Spanish-language speech

Eugenio Massolo

Ingeniería de Sonido, Universidad Nacional de Tres de Febrero
ugemassolo@gmail.com

Abstract - *This study seeks to analyze the performance of objective audio quality prediction metrics on Spanish-speaking signals, focusing particularly on the ViSQOL and CDPAM algorithms. With this objective, a MOS test of perception of audio quality was carried out as a subjective evaluation, whose results were correlated withthe values obtained by the algorithms under analysis. The results obtained indicate a high correlation between the assessment by the listeners and the estimation made by the objective metrics, although the fact that the correlation on signals in Spanish is higher than on signals in English suggests that there are differences in the methods of analysis with respect to those of the previous literature.*

1. INTRODUCTION

Comparison and evaluation of audio quality are a simple task for the human brain which can perceive very specific details and distortions. But this is a process that requires time and effort, so it is not a viable methodology for constantly evaluating signal processing algorithms that require this kind of validation. That is why automated metrics are used to predict human judgment of these tasks, which usually consist of complex and intricate handcrafted systems based on different mathematical and psychoacoustic criteria.

Over time, different metrics have been developed based on human assessment studies, e.g. PESQ [1], POLQA [2], and ViSQOL [3]. Although these algorithms widely used in the field of telecommunications are easy to compute and save a lot of time and money, they present a poor and inconsistent correlation concerning the subjective audio evaluation of people, in addition to being unstable even for small perturbations and sensitive to transformations that are imperceptible to human hearing [4].

In the search for new solutions, and with the appearance and massification of neural networks algorithms thanks to an exponential increase in data processing power (and the ability to acquire it), some audio assessment prediction models have been developed based on deep learning techniques, which make use of evaluations carried out by humans to "learn" the way in which we perceive the sound differences between signals. This implies the need for large databases of subjective evaluations about audio quality in voice signals, and the efficiency of the prediction rests on the structure of the system proposed by each developer.

One of the state-of-the-art studies in this field is the model developed by Manocha et al. [5] (with further improvement in the successor model [6]), which consists of a new perceptual audio metric based on just noticeable differences (JNDs) - the minimal change at which a difference is perceived. Results of the work indicated that the learned metric is well-calibrated with human judgments, showing a high correlation with three existing Mean Opinion Score (MOS) datasets, surpassing the performance of the baseline methods (PESQ, MSE and ViSQOL) in most scenarios.

The MOS scores datasets used were derived from different subjective evaluations of audio quality made on signals modified by different processing such as denoising, dereverberation, and even intentionally altered signals by equalization, compression and the addition of white noise.

But the human assessment of audio quality is directly affected by the language in which speech signals are heard due to the variation and distribution of vowels and consonants in different languages. This can be particularly critical for metrics generated by neural networks since the results generated by each model are biased by the data used to train it. In the case of the metrics proposed in [5] and [6], both models were trained from a set of subjective evaluations carried out with audios of voices in English. In addition to this, the MOS tests with which the model was evaluated were based on voices in English for the most part.

The present study aims to perform an analysis of the correlation between some of the different audio quality prediction metrics mentioned (particularly ViSQOL[3] and CDPAM[6]) and the subjective evaluation of speech audio quality carried out by Spanish-speaking people with speech signals in their native language.

2. METHODS

As mentioned above, the objective of the study is to find the correlation between objective metrics (ViSQOL and CDPAM) and the subjective perception of quality by Spanish-speaking people. For this purpose, a certain set of speech signals was analyzed by calculating the objective audio quality scores as well as through a subjective MOS evaluation.

The methodology in this work follows a similar experimental design and performance evaluation to Hines et al. [7]. At the end of the objective and subjective evaluations, the Pearson and the Spearman correlation coefficient were calculated between the results of the predictive metrics and the MOS tests carried out.

2.1. Test Stimuli

The set of test stimuli are made up of speech signals obtained from two Argentinian announcer voice sample web pages: Locutores [8] and Locutores Profesionales [9]. The speech samples consist of two male and two female speakers, two sentences each, making up a total of 8 signals as recommended by the ITU-T P.863 standard [10].

These signals were processed to generate a data set with different levels of audio quality. The processes carried out were the following:

- Speech denoising of previously noise-added stimuli (at 10 dB of signal-to-noise ratio). The Noise sounds were obtained from UrbanSound8K [11] dataset and consisted of vehicular traffic noise. The denoising process was carried out with RX software [12].
- Dereverberation of previously reverberated stimuli. The reverberation process was made convolving the original signals with a big room's impulse response, and the dereverberation process was done with RX software [12].
- MP3 encoding at 16 kbps [13].

Both the noise and reverberation added was made by the authors. By way of recapitulation, the subgroups of signalsare presented in Table 1.

Table 1: Stimuli dataset description.

Stimulus	Amount
Original speech signals	8
Denoised speech signals	8
Dereverberated speech signals	8
MP3 encoded speech signals	8
Total	**32**

A normalization at -20 LUFS of loudness between the signals of the dataset was carried out so as not to generate a bias in the perception of quality.

2.2. Objective evaluation

The objective evaluation consisted of processing each signal of the data set through the following audio quality perception estimation algorithms:

- **ViSQOL [3]:** Algorithm developed by Google. Based on a distance metric called the Neurogram Similarity Index Measure (NSIM), which is mapped into a MOS scale.
- **CDPAM [6]:** A deep learning based objective metric, trained over large sets of Just Noticeable Difference (JND) subjective tests of degradated signals.

The implementation of these metrics and the subsequent analysis of the results were carried out using a Python script. Both algorithms are considered intrusive methods since they require both the original and the degraded signals to reach a result.

In order to satisfy the operating conditions of the metric implementations, all signals were pre-processed to maintain the same 16-bit PCM resolution. Regarding the sampling frequency, the signals were resampled at 16 kHz for the ViSQOL algorithm and 22050 Hz for the calculation of the CDPAM metric, although the bandwidth was limited to the 8 kHz signals to maintain the same frequency content in both cases.

In the case of the CDPAM algorithm, it was necessary to convert the output scores that originally go from 1 (bad score) to 0 (good score), to a MOS scale whose lowest score is 1 and its maximum score is 5, in order to follow the guidelines of the ITU-T standard.

2.3. Subjective evaluation

The subjective evaluation was a mean opinion score (MOS) test in which listeners were asked to evaluate the quality of the sound stimuli presented to them, using a 5-point scale as suggested by the guidelines of the ITU-T standard. P.863.

According to the ITU-T recommendations [10], a minimum of 32 test responses was required to make comparisons between the algorithms., only requesting that they be native Spanish-speakers.

2.3.1. Test procedure

To carry out the subjective test remotely, the Google Forms platform was used, both to provide the stimuli and to collect the evaluations of the participants.

Listeners were asked to take the assessment with headphones regardless of their type or quality. They were also asked to do it from a computer or cell phone located in a quiet environment, free from visual and sound distractions. Then, a brief explanation of the objective of the test was provided. Before the start of the test, one of the dataset stimuli

was presented as a calibration signal and listeners were asked to set the volume of their sound system so that they were able to hear clearly.

A total of 32 stimuli were presented, for which listeners were asked to rate them through the question “Como calificaría la calidad sonora del audio en una escala del 1 al 5?”, which can be translated as "How would you rate the sound quality of the audio on a scale between 1 and 5?". In addition, Table 2 was provided, which establishes the connotation of each of the possible scale values.

Table 2: MOS Scale of signal quality.

Rating	Quality [English]	Quality [Spanish]
5	Excellent	Excelente
4	Good	Buena
3	Fair	Aceptable
2	Poor	Mediocre
1	Bad	Mala

At the end of the test, the listeners were required to complete a survey detailing their age and whether they had had previous experience with listening tests.

The MOS scores obtained for each signal were averaged over the total number of subjects participating in the test, thus obtaining a single MOS value for each signal.

3. RESULTS

3.1. Subjective evaluation

A total of 64 listeners participated in the test, discarding only two of them for having hearing problems and ruling out another two for invalidly answering the test. Therefore, 60 responses were obtained for each of the 32 audio signals of the test.

From these data, the Cronbach's Alpha coefficient was calculated in order to evaluate the internal consistency and reliability of the scale used, obtaining a score of 0.99. This score is optimal and indicates a correct use of the MOS scale by the listeners.

On the other hand, the normality of the distribution of scores for each of the subjective evaluation signals was evaluated using the Kolmogorov-Smirnov test. A $p < 0.05$ was obtained in every case, therefore normality could not be confirmed for any audio.

The MOS scores obtained for each signal were averaged over the total number of subjects participating in the test, thus obtaining a single MOS value for each signal.

3.2. Correlation between objective metrics and subjective evaluation

The scores obtained from the objective metrics were contrasted with the responses obtained from the subjective MOS test.

Figure 1 shows the contrast between the subjective evaluation of audio quality and the estimation made by the ViSQOL algorithm, while Figure 2 shows the same contrast regarding the CDPAM algorithm.

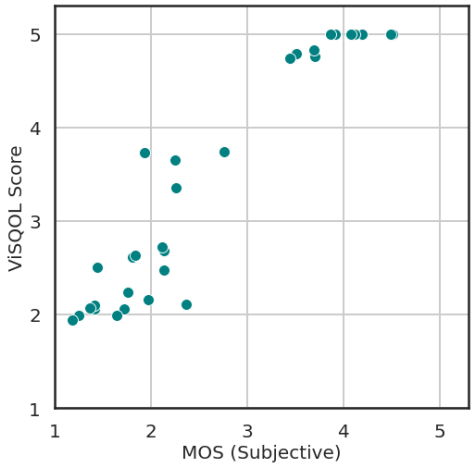

Figure 1: Subjective assessment vs ViSQOL quality estimation.

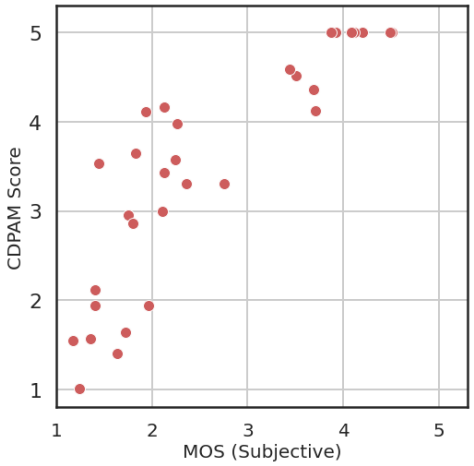

Figure 2: Subjective assessment vs CDPAM quality estimation.

In both cases, a monotonicity trend and a certain correlation between the subjective score and the estimate made by the metrics can be observed. To quantify the degree of correlation, the Pearson and Spearman coefficients are calculated for both cases. Since the normality of the distribution of the data could not be confirmed, the Pearson coefficient would not be suitable to represent the correlation, but it is presented anyway for the purpose of comparison with the previous and future literature.

Table 3 presents the results corresponding to said correlation coefficients, with a significance level of 0.01 in both cases.

Table 3: Correlation between subjective MOS and objective metrics.

Algorithm	Pearson coef.	Spearman coef.
ViSQOL	0.95	0.93
CDPAM	0.87	0.91

A high level of correlation can be observed between both algorithms and the assessment made by the listeners in the subjective test. In addition to the high degree of correlation, the performance of both metrics in predicting the audio quality assessment is similar, particularly if the Spearman coefficient is used for the comparison.

4. DISCUSSION

Despite the high degree of correlation, it is important to highlight the rarity of the fact that the heats of correlation between objective metrics and the subjective evaluation on Spanish-speaking signals are greater than the correlation values on English signals presented by the developers of the algorithms.

The correlation coefficients presented by the developers of the CDPAM metric [6], which compares the performance of their own algorithm with ViSQOL (among others) on English-speaking signals, are presented in Table 4, together with the results obtained in this investigation.

Table 4: Correlation values over English-speaking and Spanish-speaking signals.

	Spearman coefficient		
Language	**CDPAM**	**ViSQOL**	**p-value**
English	0.75	0.40	< **0.05**
Spanish	0.91	0.93	< **0.01**

The main reason to doubt the results is that both algorithms were designed to operate with English-speaking signals, therefore their best performance should be on those signals.

The case of the CDPAM algorithm is even more striking, since its operation based on neural networks is strongly dependent on the information used for its training, in this case being English speaking signals.

Although it may be the case that the correlation on signals in Spanish is higher than that obtained on the original language, the large difference in the degree of correlation between languages (0.16 for CDPAM and 0.53 in the case of ViSQOL) suggests that the procedure carried out in this research differs somewhat from that used in the previous literature [6].

Given that the implementation of the algorithms was carried out according to its documentation, both in terms of the duration of the signal and the sampling frequency, the subjective test is considered as the potential factor of incidence on the results. The major difference between the evaluations made by previous studies [6] [7] and the current one lies in the amount of stimuli used in the test (more than 5 times higher) and the variety of processing used to degrade speech signals. This directly affects the different levels of audio quality presented by the evaluation, since a greater variety of degradation generates differences that are less discernible by listeners, while in the opposite case, participants may notice the different "subgroups" of audio quality, leading them to make more consistent assessments.

This limitation in terms of the amount of stimuli is given by the impossibility of making the listeners evaluate too many audio signals, since this leads to auditory fatigue and therefore to the invalidation of their evaluations. Although different signal datasets could be made to be evaluated by different groups of participants, many more listeners would be required to perform the test.

Regarding the insufficiency in the variety of degrees of distortion of the signal, it is due to the inability to access the same degradation techniques that were used in previous studies, which would allow a much more representative comparison of the performance of the algorithms.

Leaving aside the difference in absolute terms in the degree of correlation, a greater change in the correlation level of the VisQOL algorithm with respect to that of CDPAM can be observed when applied to Spanish language signals. This can be attributed to the fact that, as mentioned above, the CDPAM algorithm is based on neural networks and requires training from audio signals that in this case were from the English language, so its performance on signals that differ from the "known" ones implies a limitation in their performance. While on the other hand, the ViSQOL algorithm works by means of mathematical calculations on the temporal and frequency spectrum of the signal, which is more suitable for its application in any type of speech signal, regardless of its language.

5. CONCLUSIONS

A subjective audio quality perception test was successfully carried out, highlighting the good performance and intuitive use of the BeaqleJS framework, whose results based on the MOS scale showed excellent consistency in its use.

However, the correlation obtained between the subjective assessment and the estimation made by the objective metrics yielded results that can be classified as "too optimistic", since the fact that the algorithms perform better in Spanish-speaking signals than in

signals speaking in English (the language for which they were designed) is hard to believe.

The objective of future studies is to create a data set using the same speech signal degradation techniques that were implemented in the studies carried out by Manocha et al. [6] and thus obtain results that are suitable for a direct comparison, and therefore more appropriate to draw conslusions regarding the performance of the objective metrics.

REFERENCES

[1] A. W. Rix and J. G. Beerends et al., “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), vol. 2. IEEE, 2001, pp. 749–752.

[2] J. G. Beerends, C. Schmidmer et al., “Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part itemporal alignment,” Journal of the Audio Engineering Society, vol. 61, no. 6, pp. 366 – 384, 2013.

[3] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, “Visqol: an objective speech quality model,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2015, no. 1, pp. 1–18, 2015.

[4] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, “Robustness of speech quality metrics to background noise and network degradations: Comparing visqol, pesq and polqa,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 3697–3701.

[5] P. Manocha, A. Finkelstein, et al., “A differentiable perceptual audio metric learned from just noticeable differences,” in Interspeech, 2020.

[6] P. Manocha, N. Bryan, G. Mystore, et al., “CDPAM: Contrastive learning for perceptual audio similarity,” in Interspeech, 2020.

[7] A. Hines, J. Skoglund, A. Kokaram and N. Harte, "Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 3697-3701, doi: 10.1109/ICASSP.2013.6638348.

[8] Locutores. www.locutores.net. (Last viewed on 17/10/2021).

[9] Locutores Profesionales. www.locutoresprofesionales.net (Last viewed on 17/10/2021).

[10] ITU, “Perceptual objective listening quality assessment,” Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.863, 2011.

[11] J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research", 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.

[12] RX9, The industry standard for audio repair. https://www.izotope.com/en/products/rx.html (Last viewed on 20/11/2021)

[13] Audioconvert. https://www.aconvert.com/audio/ (Last viewed on 17/10/2021).