Scientific Research Publishing

Open Journal of Applied Sciences, 2021, 11, 1324-1332
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Impact of Languages and Accent on Perceived Speech Quality Predicted by Perceptual Evaluation of Speech Quality (PESQ) and Perceptual Objective Listening Quality Assessment (POLQA): Case of Moore, Dioula, French and English

Daouda Konane[1], Sibiri Tiemounou[2], Wend Yam Serge Boris Ouedraogo[1]

[1]Laboratoire de Matériaux et Environnement, Université Joseph KI-ZERBO, Ouagadougou, Burkina Faso
[2]Institut du Génie Informatique et Télécommunications, Ecole Polytechnique de Ouagadougou, Ouagadougou, Burkina Faso
Email: dkonane87@gmail.com, sibiri.tiemounou@gmail.com, ouedboris@gmail.com

How to cite this paper: Konane, D., Tiemounou, S. and Ouedraogo, W.Y.S.B. (2021) Impact of Languages and Accent on Perceived Speech Quality Predicted by Perceptual Evaluation of Speech Quality (PESQ) and Perceptual Objective Listening Quality Assessment (POLQA): Case of Moore, Dioula, French and English. *Open Journal of Applied Sciences*, **11**, 1324-1332.
https://doi.org/10.4236/ojapps.2021.1112100

Received: November 2, 2021
Accepted: December 27, 2021
Published: December 30, 2021

Abstract

Perceptual Objective Listening Quality Assessment (POLQA) and Perceptual Evaluation of Speech Quality (PESQ) are commonly used objective standards for evaluating speech quality. These methods were developed and trained on native speakers' speech sequences of some western languages. One can then wonder how these methods perform if they are applied to other languages or if the speaker is non-native. This paper deals with the evaluation of PESQ and POLQA on languages that were not been considered when setting up these methods, with emphasis on Moore and Dioula, two local languages of Burkina Faso. Another aspect is the evaluation of these two methods in the case of non-native speakers. For this purpose, in the one hand, the Mean Opinion Score-Listening Quality Objective (MOS-LQO) of PESQ and POLQA, computed for Moore and Dioula, are compared to those of French and English. On the second hand, the MOS-LQO scores of French and English are compared for native and non-native speakers, to evaluate the effect of the accent of speakers.

Keywords

Speech Quality, PESQ, POLQA, Language, Accent

1. Introduction

Standards for assessing perceived speech quality can be divided into two main

groups: subjective and objective methods.

Subjective methods, also called subjective testing, consist of a set of tests in which participants judge the speech quality as they perceive it on a defined quality scale [1] [2]. The scores from this test are used to calculate an average score called Mean Opinion Score (MOS). This approach is the most suitable for assessing speech quality. However, it can be expensive and time-consuming to implement. Objective methods, also called objective models, aim to automatically predict the perceived speech quality as it would be obtained in a formal subjective test. The predicted score, called MOS-LQO (Mean Opinion Score—Listening Quality Objective), is obtained by comparing a degraded signal and its original version [1] [3] [4] [5]. The well-known and most widely used objective standards, by telecommunications operators, are POLQA (ITU-T Standard P.863 [6]) and PESQ (ITU-T standard P.862 [5]). POLQA can be used to evaluate speech quality in Narrow Band ([300 Hz; 3400 Hz]), Wide Band ([50 Hz; 7000 Hz]), and Super Wide Band (50 Hz; 14,000 Hz). Regarding PESQ, it only operates on narrow and wide bands.

One should note that these two speech quality measurement models were developed and trained on native speakers' speech sequences of some western languages. For example, 11 languages were used for POLQA [7] [8] namely: English, British English, Chinese (Mandarin), Czech, Dutch, French, German, Swiss, German, Italian, Japanese, Swedish. However, these standards are used in several countries, including Burkina Faso, by telecommunications regulatory authorities and by telephone operators to evaluate the speech quality transmitted in phones networks. One then wonder how these methods perform if they are applied to other languages, or if the speaker is non-native. Several authors have already carried out work on similar issues. F. Ben Ali *et al.*, [9] investigated the dependency on the language and objective quality assessment models. By using an important measurements database, they mapped the scores of SwissQual's speech quality algorithm for Listening Quality (Squad-LQ) and PESQ, for the languages French, English, and Arabic. They concluded that PESQ and Squad-LQ do not score these three languages in the same way. By working on English and Igbo (a West African tonal language), D. U. Ebem *et al.*, [10] showed that the MOS-LQO scores predicted by POLQA for Igbo, seem to be overestimated compared to the MOS scores given by Igbo listeners.

No previous scientific work has focused on the evaluation of PESQ and POSQA on the local languages of Burkina Faso.

This work compare the MOS-LQO scores of PESQ and POLQA for Moore and Dioula (two local languages of Burkina Faso), with those of French and English. The speech sequences of these four languages, considered to compute the MOS-LQO scores, come from native and non-native speakers.

The aims are, on the one hand, to evaluate the influence of the languages on the MOS-LQO scores and on the other hand, to evaluate the impact of a native and non-native speaker. It should be noted that in Burkina Faso, telephone communications are carried out in a narrow band. Therefore, the POLQA and

PESQ models will be evaluated in this band.

The continuation of the document is organized into three parts. The first part presents the process used to record the speech sequences. The second part shows the obtained results and discussions, while the third part derived conclusions and perspectives.

2. Methodology for Speech Sequences Database Construction

2.1. Reference Speech Signals Database Construction

To constitute the database of reference speech signals, let considered the four languages: Moore, Dioula, French, and English. For each language, two double sentences are pronounced by four speakers (two men and two women, all from Burkina Faso). A record thirty-two original speech signals samples of (*i.e.* eight per language) is performed. Each speech signal is 8 seconds in duration and is chosen to contain several sounds of the considered language [4] [11] [12] [13]. In addition, 8 speech signals from native French and 8 speech signals from native English were extracted from the database used by S. Tiemounou for testing and validating the POLQA and Diagnostic Instrumental Assessment of Listening quality (DIAL) standards [7] [8].

A total set of 48 reference speech signals were constructed. These speech signals are sampled at 48 kHz and quantized on 16 bits. However, to simulate narrowband communication, these signals were down-sampled to 8 kHz and then degraded by adding different nature's defects as described in the following section.

2.2. Degraded Speech Signals Database Construction

In the purpose to simulate the defaults perceived during phone calls, different degradation conditions have been considered, as described in Table 1. Leman *et al.* [12] [13] and Tiemounou *et al.* [11] have shown that the noises perceived during narrowband and super-wideband phone calls can be subdivided into three families, among which environmental and breathing noises are the most representative. One can choose babble noise to model the noise of the environment and a random pink noise for the breath one.

To cover a wide range of perceived noise levels, let chose five Signal to Noise Ratio (SNR) values (0, 10, 20, 30, and 40 dB) for each kind of noise. This leads to a total of 10 degradation conditions due to noise. In addition, degradations relative to the variation of the sound level (loudness) (−5; −10; −15 and −20 dB SPL)

Table 1. Summary of degradation conditions.

Type of degradation	Details
Babble noise (non-stationary noise)	RSB = 0, 10, 20, 30, 40 dB
Pink noise (stationary noise)	RSB = 0, 10, 20, 30, 40 dB
Sound level Loudness	Level = −5, −10, −15, −20 dB SPL
No degradation	

as welle as no degradation case (corresponding to the reference signal) are considered. A total of 15 degradation conditions were performed.

Figure 1 describes the degraded speech signals generation process. It should be noted that the same approach described in [7] is adopted. The database was constructed in such a way as to simulate a narrowband communication from the 48 reference speech samples and the 2 background noises. First, the speech signal is down-sampled to 8 kHz and filtered [7] [13] to obtain a narrowband signal (from 300 to 3400 Hz). Then, the resulting signal is equalized to −26 dBov according to ITU-T P.56 [14]. For degradation due to noise, the reference speech and noise signals are mixed (with different SNR) to obtain the degraded signals. In addition, the resulting signal level is again equalized and then coded and decoded using the G 711 code [15], one of the most widely used codecs by cell phone operators in Burkina Faso. This process leads to the degraded signal. For degradation due to the sound level, there is no mixing with noise, but the degradation is performed during the sound level equalization step.

Then, 720 degraded speech signal samples were generated.

3. Results and Discussions

3.1. Impact of Language on MOS-LQO Scores of PESQ and POLQA

This section presents the evaluation results of PESQ and POLQA on the degraded signals generated in Section 2. Figure 2 shows the MOS-LQO scores of PESQ and POLQA for the four languages (Moore, Dioula, Native French, and

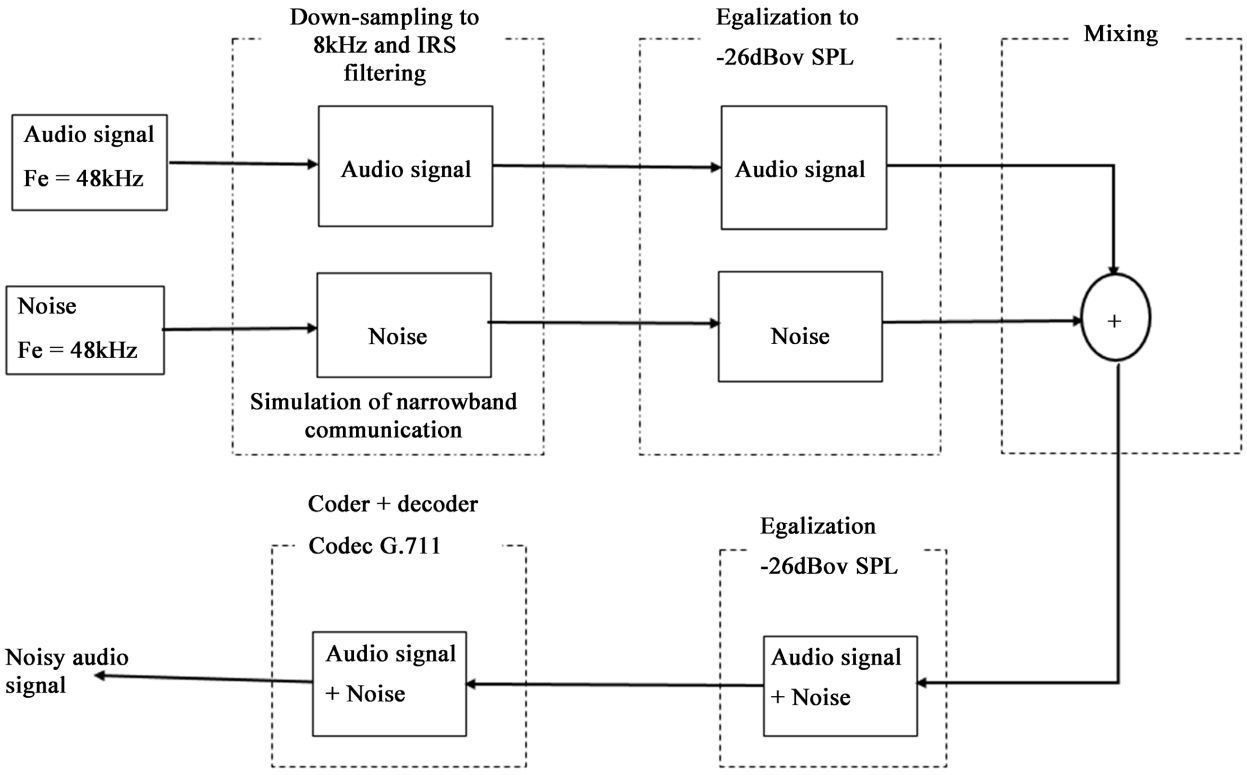

Figure 1. Block diagram for generating degraded signals.

Native English) obtained by Monte Carlo simulation [16].

One can see in Figure 2, that for Moore and Dioula the MOS-LQO scores of PESQ are larger than those of native French and native English. On the other hand, for POLQA, no language seems to emerge.

In the following paragraph, a statistical analysis of the MOS-LQO scores of PESQ and POLQA is performed, to validate the results obtained above. To measure the impact of the language on the MOS-LQO scores of PESQ and POLQA, let used the ANalysis Of VAriance (ANOVA) method [17]. It is an inferential statistical method that tests whether the means of several groups are significantly different. The statistical hypotheses are the following:

- H0 (or null hypothesis): all means are equal;
- H1 (or alternative hypothesis): all means are not equal.

To validate the null hypothesis, a significance threshold (denoted alpha) must be specified. The ANOVA test provides 2 main statistical values:

- F: it corresponds to the ratio of the variation between the means of the samples and the variation within the samples;
- p-value: probability associated with the F statistic.

Thus if the p-value is lower than alpha then the null hypothesis is rejected. Therefore, the means are statistically different. Otherwise, one cannot make a decision.

Figure 3 shows the distribution (boxplot) of MOS-LQO scores of PESQ and POLQA for the four languages (Moore, Dioula, native French, native English), and Table 2 displays the ANOVA results applied to the four languages.

As one can see in Table 2, the p-value for the POLQA model is very large (0.98), so the null hypothesis cannot be rejected. Therefore, one can conclude that for the POLQA model, language does not seem to have an impact on the

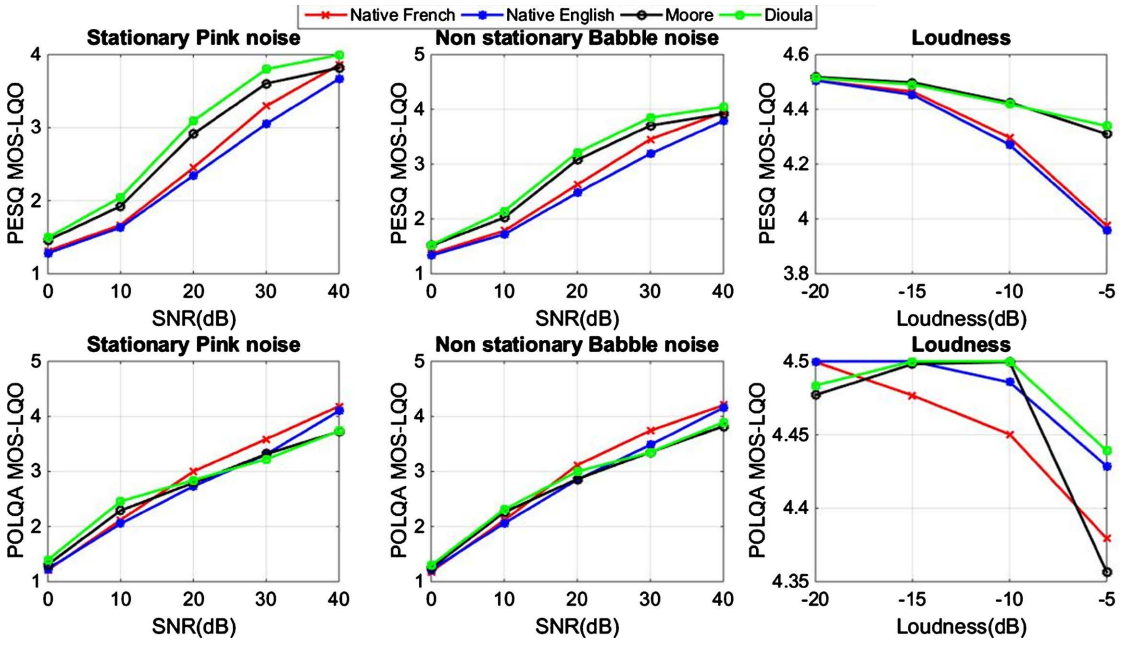

Figure 2. MOS-LQO scores of PESQ and POLQA for the four languages (Moore, Dioula, native French, native English).

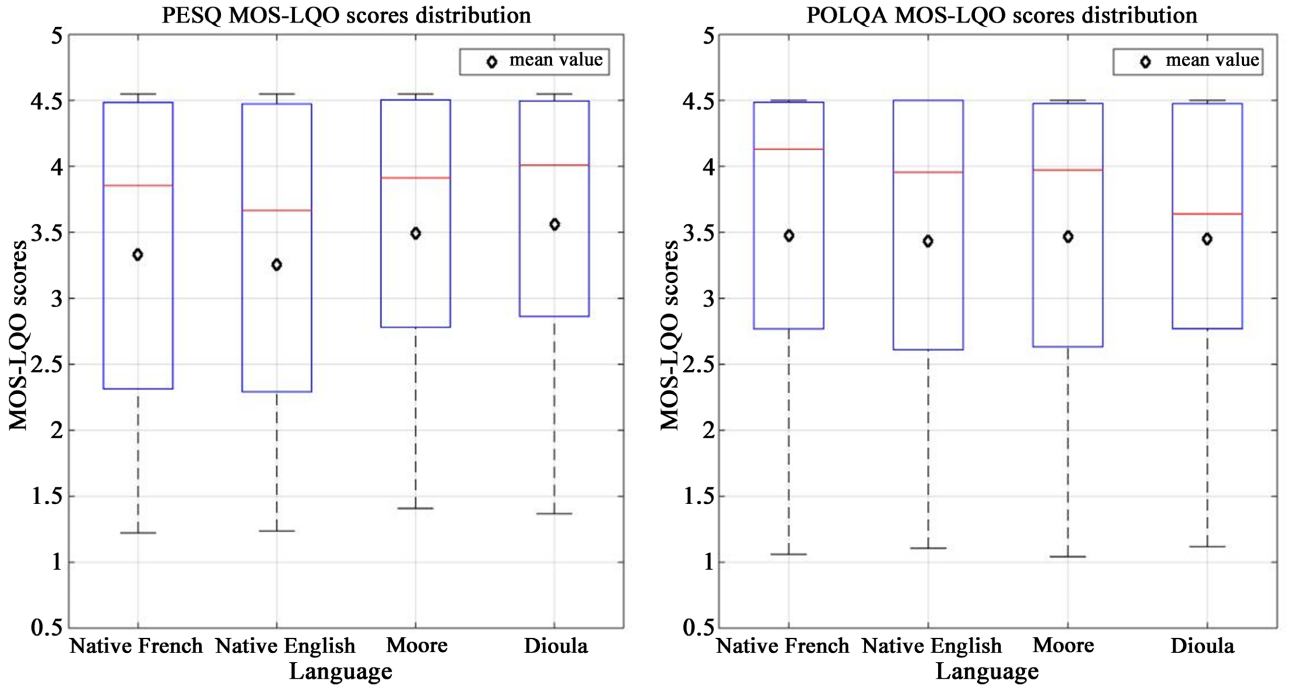

Figure 3. Boxplot of MOS-LQO scores for the four languages.

Table 2. ANOVA results.

	Average value of MOS-LQO				ANOVA Statistics	
	Native French	Native English	Moore	Dioula	F Statistics	p-value
POLQA	3.48	3.43	3.47	3.45	0.05	0.98
PESQ	3.33	3.25	3.49	3.56	2.1	0.09

predicted speech quality. On the other hand, for PESQ, the null hypothesis is rejected for a significance threshold of 0.1. Thus, language does have an impact on the voice quality predicted by PESQ.

3.2. Impact of Accent on MOS-LQO Scores of PESQ and POLQA

Figure 4 compares the MOS-LQO scores of PESQ and POLQA for native and non-native speakers, using French and English. The native speaker's speech signals come from [7], while the non-native speech sequences are recorded from Burkina Faso speakers. Figure 3 also shows that for PESQ, the MOS-LQO scores of native speakers are larger than those of non-native speakers. However, these results are not obvious for POLQA.

To confirm the previous result, a one-sided Student's test (a variant of the ANOVA for the case of 2 groups) is performed. The distribution of the MOS-LQO scores of these two groups are presented in Figure 5 and Table 3 presents the average MOS-LQO scores for the two groups (native and non-native), as well as the Student's test statistics.

Table 3 presents the average MOS-LQO scores for the two groups (native and non-native), as well as the results of Student's test statistics.

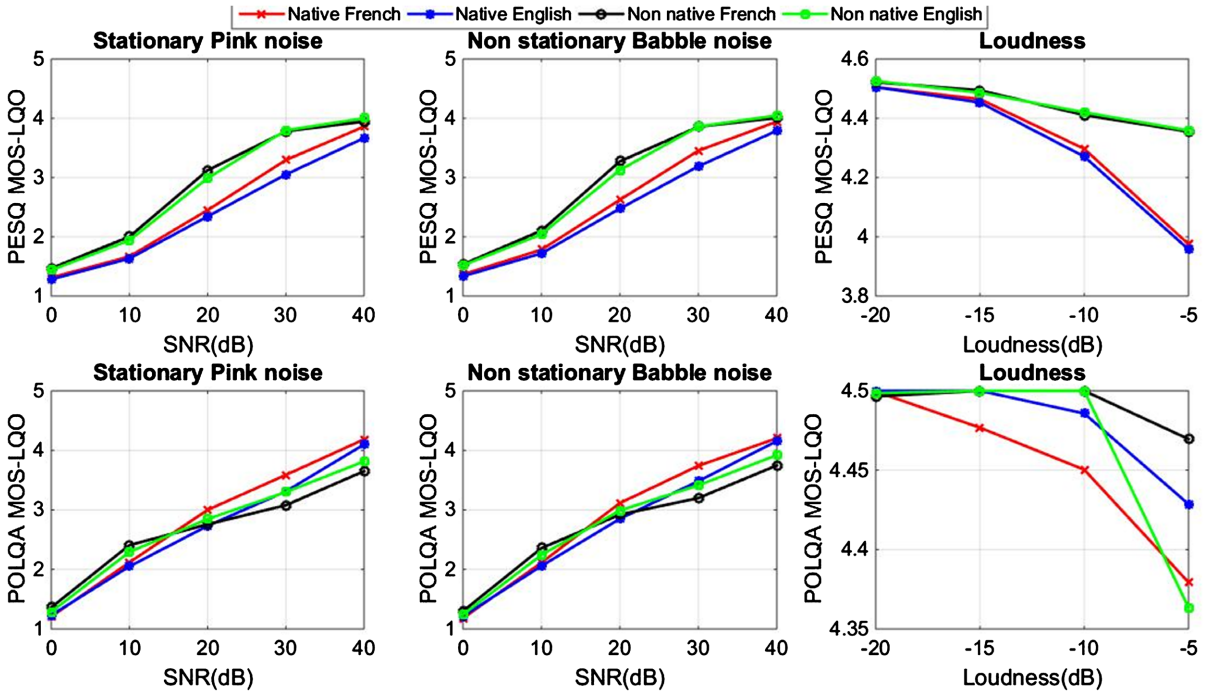

Figure 4. MOS-LQO scores of PESQ and POLQA for native and non-native.

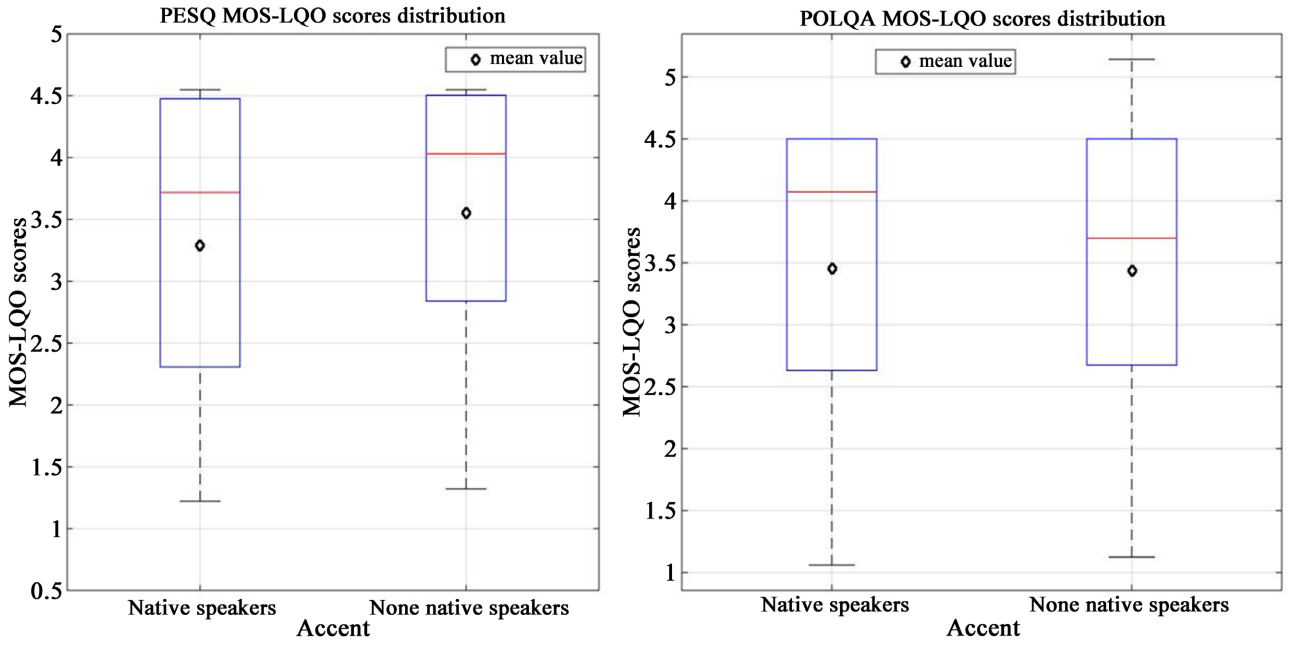

Figure 5. Boxplot of MOS-LQO scores for natives and non-natives speakers

Table 3. Student test results.

	Average value of MOS-LQO		Student test statistics	
	non-native speaker	native speaker	T statistics	p-value
POLQA	3.45	3.44	0.59	0.28
PESQ	3.29	3.55	2.62	0.0045

Here again, for the PESQ standard, the null hypothesis can be rejected for a significance threshold of 0.01. While for POLQA the null hypothesis cannot be rejected. However, the relatively low value of the p-value requires further study.

The difference of performance observed between POLQA and PESQ could be explained by some limitation of PESQ. Indeed, it was shown in [18] that PESQ performs worst for some codecs, like Enhanced Variable Rate Codec (EVRC) family codecs [19], VoIP systems [20] and for wideband signal. The degradation caused by all these system may concern the signal spectral content. One can think that in case of degradation due to noise, the spectral content for Moore, Dioula or non native speaker does not modify in the same way than those of French or english native speaker. This, can affect PESQ performance. As mention in ITU-T P.863 [3], POLQA was developped in order to overcome PESQ limitations, in particular any variation of the signal spectral content, such as Spectral flatness, Strong variations of the Disturbance Density over time indicators. This could explain why POLQA performance is insensitive to language or accent. Nevertheless, futer work will focus on the study of spectral contents of the four languages.

4. Conclusions and Perspectives

This paper evaluates the impact of language and accent on the speech quality predicted by the PESQ and POLQA standards. First of all, the effect of language is evaluated by comparing the MOS-LQO scores of PESQ and POLQA obtained for the language Moore and Dioula (two languages of Burkina Faso), as well as for French and English. In a second step, the effect of accent is evaluated by comparing the MOS-LQO scores of the two standards on speech signals of French and English from native and non-native speakers. The results show that language and accent significantly impact the perceived speech quality predicted by PESQ, but seem to have no significant effect on POLQA.

Future work include experiments with a cell phone operator of Burkina Faso and comparison of the MOS-LQO scores predicted by the PESQ and POLQA models with the subjective MOS scores delivered by the listeners.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] ITU-T (1996) Method for Subjective Determination of Transmission Quality, ITU-T Recommendation P.800. https://www.itu.int/rec/T-REC-P.800-199608-I

[2] ITU-T (2016) Mean Opinion Score Interpretation and Reporting, ITU-T Recommendation P.800.2. https://www.itu.int/rec/T-REC-P.800.2-201607-I/en

[3] ITU-T (2018) Perceptual Objective Listening Quality Prediction, ITU-T Recommendation P.863. https://www.itu.int/rec/T-REC-P.863

[4] ITU-T (2007) Application Guide for Objective Quality Measurement Based on Recommendations P.862, P.862.1 and P.862.2, ITU-T Recommendation P.862.3.

[5] ITU-T (2001) Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, Rec. P.862. https://www.semanticscholar.org/paper/ITU-T-Rec.-P.862.3-(11%2F2007)-Application-guidefor/0dbc078520210285a77949f4d4b9271055be2924

[6] Rix, A.W., Beerends, J.G., Hollier, M.P. and Hekstra, A.P. (2001) Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs. 2001 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **2**, 749-752.

[7] Tiemounou, S. (2014) Développement d'une méthode de diagnostic technique des dégradations de qualité vocale perçue des communications téléphoniques à partir d'une analyse du signal de parole, Thèse soutenue à Rennes le 17 février 2014.

[8] ITU-T (2011) Perceptual Objective Listening Quality Assessment, ITU-T Recommendation P.863. https://www.itu.int/rec/T-REC-P.863/fr

[9] Ben Ali, F., Djaziri Larbi, S., Jaïdane, M. and Ridane, K. (2009) Experimental Mappings and Validation of the Dependence on the Language of Objective Speech Quality Scores in Actual GSM Network Conditions. 2009 17*th European Signal Processing Conference*, Glasgow, Scotland, August 24-28, 2009, pp. 2534-2538.

[10] Ebem, D.U., Beerends, J.G., Van Vugt, J., Schmidmer, C., Kooij, R.E. and Uguru, J.O. (2011) The Impact of Tone Language and Non-Native Language Listening on Measuring Speech Quality. *Journal of the Audio Engineering Society*, **59**, 647-655.

[11] Tiemounou, S., Le Bouquin Jeannès, R. and Barriac Ewert, V. (2014) Perception-Based Automatic Classification of Background Noise in Super-Wideband Telephony. *Journal of the Audio Engineering Society*, **62**, 776-781.

[12] Leman, A., Faure, J. and Parizet, E. (2008) Influence of Informational Content of Background Noise on Speech Quality Evaluation for VoIP Application. *The Journal of the Acoustical Society of America*, **123**, 471-476.

[13] Leman, A., Faure, J. and Parizet, E. (2009) A Non-Intrusive Signal-Based Model for Speech Quality Evaluation Using the Automatic Classification of Back-Ground noises. *InterSpeech* 2009, Brighton, 2009, p. 1.

[14] ITU-T (2011) Objective Measurement of Active Speech Level, ITU-T Recommendation P.56. https://www.itu.int/rec/T-REC-P.56

[15] ITU-T (1988) Pulse Code Modulation (PCM) of Voice Frequencies, Recommendation ITU-T G.711. https://www.T-REC-G.711-198811-I!!PDF-F.pdf

[16] Raychaudhuri, S. (2008) Introduction to Monte Carlo simulation. *2008 Winter Simulation Conference*, Miami, Florida, USA, December 7-10, 2008, 91-100. https://doi.org/10.1109/WSC.2008.4736059

[17] Kim, H.Y. (2014) Analysis of Variance (ANOVA) Comparing Means of More Than Two Groups. *Restorative Dentistry & Endodontics*, **39**, 74-77.

[18] Engineering Services Group (2008) PESQ Limitations for EVRC Family of Narrowband and Wideband Speech Codecs. White Paper, 2008.

[19] 3GPP2 C.S0014-D v3.0 (2010) Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, and 73 for Wideband Spread Spectrum Digital Systems. www.3gpp2.org

[20] Manjunath, T. (2009) Limitations of Perceptual Evaluation of Speech Quality on VoIP Systems. 2009 *IEEE International Symposium on Broadband Multimedia Sytems and Broadcasting*, Bilbao, Spain, 13-15 May 2009, 1-6. https://doi.org/10.1109/ISBMSB.2009.5133799