PERFORMANCE EVALUATION OF QUALITY DEGRADATION INDICATORS ON SUPER-WIDEBAND SPEECH SIGNALS

Sibiri Tiemounou[1,2,3], Régine Le Bouquin Jeannès[2,3], Vincent Barriac[1]

[1] Orange Labs - Lannion, 2 Av. Pierre Marzin, 22307 Lannion Cedex, France
[2] INSERM, U 1099, Rennes, F-35000 France
[3] Université de Rennes 1, LTSI, Rennes, F-35000, France
{sibiri.tiemounou, vincent.barriac}@orange.com, regine.le-bouquin-jeannes@univ-rennes1.fr

ABSTRACT

This paper presents the performance of quality degradation indicators to be used in the context of super-wideband (50-14000 Hz) telephony. After an overview of these indicators, two analyses are undertaken: the first one considers conditions containing a single degradation and the second one considers conditions comprising several degradations at the same time, reflecting more realistic communications. This study highlights the major role of some indicators, and particularly those designed for quantifying the perceived additive noise, the frequency-response distortion and also the speech level. We show that these indicators are robust to multiple types of degradations and reveal relevant for advanced diagnosis of telecommunication systems.

Index Terms— perceptual dimensions, quality degradation indicators, super-wideband (SWB), voice quality assessment, objective models

1. INTRODUCTION

During speech transmission over modern telecommunication systems, various disturbances and deformations occur, with a non negligible impact on perceived voice quality. In order to fix these issues, it is important to diagnose them first. To this end, it has been proposed in [1] to divide the voice quality degradation in three dimensions, assumed to be mutually orthogonal: time degradations, frequency-response degradations and additive-noise degradations. In order to get more detailed information about these families of degradations, several quality degradation indicators have been proposed [1-4]. For instance, in [2], Leman *et al.* proposed parametric and hybrid indicators to quantify the impact of time degradations and signal-based indicators to diagnose the effect of background or circuit noise on speech quality.

Furthermore, most of recent speech listening quality evaluation models, particularly the recent ITU-T P.863 standard [3] (or also, for instance, the model proposed in [4]), have included, beside the three degradation dimensions described above, quality degradation indicators that quantify speech level deviations and thus have integrated them into a specific fourth dimension. This final set of four perceptual dimensions is assumed in what follows to cover the whole speech quality space present in modern telecommunication networks and services.

The work presented here is only focused on the performance evaluation of signal-based quality degradation indicators in the context of super-wideband speech signals. In our preceding study realized in [5], we presented a visual comparison between quality degradation indicators that we identified in [3] and [4]. However, this approach gave only an overview on their characteristics (their monotony, in particular), not a full assessment of their performance. To fill this lack, in this paper, we adopted some performance criteria similar to the approach proposed in [6]. It is important to note that our goal is to select and propose the most relevant and reliable quality degradation indicators, both in terms of accuracy and robustness.

This paper is organized as follows: Section 2 highlights the 4 perceptual dimensions and their sub-dimensions as found in literature. For each dimension, the quality degradation indicators we selected, or built, are described in Section 3. After a detail description of the criteria to assess the performance of these quality degradation indicators in Section 4 and the way we applied them on a super wide-band speech database in Section 5, results are presented and discussed in Section 6, before drawing conclusions in Section 7.

2. PERCEPTUAL DIMENSIONS

We assumed that the degradations dimensions introduced above were related to perceptual dimensions as indicated in [7]. With the help of a series of auditory experiments and multidimensional analysis (MDA) [8] performed on speech samples processed by transmission systems, three mutually orthogonal perceptual dimensions have been identified:

- Directness/Frequency Content (DFC) or Coloration: this dimension is linked to frequency-response degradations due to band-pass filtering, electro-acoustic properties of terminal equipment, and room acoustics;
- Continuity: it reflects all effects of time-varying distortions due to packet loss, bit errors or signal processing;

- Noisiness: this dimension describes the perceived additive-noise degradations due to background noise or circuit noise.

Concerning the fourth dimension, namely Loudness, considered in this work, several studies (e.g. [9]) introduced the listening level as an additional feature of the integral speech quality.

Recent studies have shown that Directness/Frequency Content, Continuity and Noisiness dimensions can be subdivided into sub-dimensions that are described hereafter. These sub-dimensions have been also revealed using auditory tests and multidimensional scaling.

2.1. Directness/Frequency Content (DFC)

This dimension is subdivided in two sub-dimensions as proposed in [10]:

(a) *Directness* or *Nearness*: sub-dimension including talking-room reflections and bandwidth limitation.

(b) *Frequency Content*: sub-dimension related to the impact of frequency response of transmission systems on speech quality.

2.2. Continuity

The studies achieved in [11] concluded that the Continuity dimension could be subdivided into three sub-dimensions:

(a) *Interruptedness*: perceived interruptions of transmitted speech (dependent on packet/frame loss rate and the concealment technique like silence insertion).

(b) *Additive artifacts*: perceived effect of frame repetition potentially generated by signal or packet processing features.

(c) *Musical noise*: sub-dimension covering time varying residual noise components due to imperfect noise reduction algorithms.

We found this subdivision more adapted to technical causes diagnosis than the one proposed by [12], more based on sound perception.

2.3. Noisiness

In [13], three sub-dimensions were proposed for this dimension. We adopted them for our study:

(a) *Speech contamination*: perception of noise-like distortions correlated with speech or lying within the (band-limited) transmitted speech spectrum.

(b) *Additive noise level*: perceived level of additive noise.

(c) *Noise coloration*: spectral shape and spectral content of noise.

3. QUALITY DEGRADATION INDICATORS

The quality degradation indicators we described hereafter per dimension were exclusively extracted from objective models, here POLQA [3] and a model proposed in [4].

3.1. Directness/Frequency Content (DFC)

(a) *Directness*: two quality degradation indicators fall in the scope of this sub-dimension. The first one is "*Erb*" (Equivalent Rectangular Bandwidth), relative to bandwidth limitation of the frequency response. It is computed from the global frequency response of the system. The second one, "*reverb*", quantifies the effect of room reverberation. It is computed from the combination of energy of the three loudest reflections.

(b) *Frequency Content*: it is characterized by the "f_c" indicator which is the central frequency of the gain of the overall transmission system.

Besides, we selected from [3] two other quality degradation indicators that address both *Directness* and *Frequency Content* sub-dimensions: the "*freq*" indicator which is used to quantify the impact of overall global frequency response distortions, similarly to the "*FRQ*" indicator used in [1], and the "*Flatness*" indicator quantifying the impact of timbre distortions and also referred as Coloration.

3.2. Continuity

From our analysis of [3], [4] and [11], we identified for the Continuity dimension five quality degradation indicators, three for *Interruptedness* and two for *Additive artifacts*:

(a) *Interruptedness*: the "r_I" indicator quantifies the rate of long level interruptions introduced in the degraded speech signal when lost frames are replaced by silence frames. It is computed from the difference between the reference and the degraded envelopes using a threshold. In addition, the "r_L" indicator represents the rate of short level interruption occurred on the degraded speech signal during speech activity periods. It is computed from the short level variations found when comparing the reference and degraded signals. The "*TimeClip*" indicator is derived from the calculation of the internal representation of speech signals, where the impact of time clipping on the perceived speech quality is modeled.

(b) *Additive artifacts*: the "r_A" indicator estimates the rate of artifact perceived on speech signal due to frame repetition. In addition, the "*frameRepeat*" indicator, derived from a comparison of the correlation of consecutive frames of the reference signal with the correlation of consecutive frames of the degraded signal, estimates severe distortions introduced by frame repetitions.

In the literature, we found no indicator to quantify the effect of musical noise.

3.3. Noisiness

Concerning this dimension, we found only quality degradation indicators for *Speech contamination* and *Additive noise level* sub-dimensions for super-wideband signals, and none for *Noise coloration*.

(a) *Speech contamination*: the "*NoS*" (Noise on Speech) indicator is used to quantify the impact of additive noise present during active speech periods.

(b) *Additive noise level*: the "*Noise*" indicator quantifies the impact of additive noise on the whole signal. It is calculated from the spectrum of the degraded signal averaged over the silent frames of the reference signal. In addition, the "*Ln*" indicator, which is the total perceived noise loudness, computed during silence periods (taking into account abrupt noise level variations), and the "*Noise contrast*" indicator, derived from the silent parts of the reference signal, allow quantifying severe noise level variations.

3.4. Loudness

Three indicators are proposed to quantify this dimension: the "*LTL*" (Long-Term Loudness) indicator for estimating the perceived loudness of the whole degraded speech signal, the "*Leq*" (Equivalent Continuous Sound Level) indicator corresponding to the mean energy of the degraded signal over all active speech frames and the "*Level*" indicator, derived from the signal level of the degraded signal, which is used to quantify severe deviations of the optimal listening level.

4. EVALUATION PRINCIPLE

The performance evaluation of quality degradation indicators is a difficult task since, in a real communication, multiple degradations often occur simultaneously. Our criteria to assess the performance of quality degradation indicators were mainly based on a study described in [6] where a Technical Cause Analysis (TCA) benchmark is proposed. A TCA indicator is defined as an indicator which should allow finding the underlying technical causes for certain types of degradation. In our study, we assumed that the quality degradation indicators could be used as TCA indicators and should respect these requirements. Mainly two requirements were considered to assess the performance such indicators:

(*4.a*) a mapped (or predicted [6]) MOS (Mean Opinion Score) derived from a TCA indicator should have a high correlation with auditory test results, preferably above 0.9, for degradations for which the indicator was designed;

(*4.b*) a TCA indicator should also have good discrimination properties, *i.e.* the corresponding mapped MOS value should be as high as possible, preferably above 3.0, for degradations for which this indicator was not designed.

In our study, for requirement (*4.a*), we chose to compute a correlation between the values of the quality indicators themselves (instead of derived MOS scores) and the corresponding auditory test results.

5. PERFORMANCE EVALUATION

To evaluate the performance of the quality degradation indicators following the principles described in Section 4, two tests were performed:

The first one (*Test 1*) consisted in estimating the performance of quality degradation indicators according to requirements (*4.a*) and (*4.b*) by testing conditions composed of only one degradation. To do so, we firstly selected 10 common anchor conditions from different SWB databases (with authorization of their owners) developed within the speech data pool of ITU-T study group 12 Question 9 set up for the POLQA benchmark (see Table 1). Except for the reference SWB signal (condition C1), each condition was only concerned by one degradation. These 10 conditions were clustered per dimension, and a total of 432 speech stimuli were taken from 4 databases including 4 languages (French, Dutch, British English and Swiss German). All conditions were represented by the same number of male speech samples, female speech samples and languages. Besides, results of auditory tests were available for these speech stimuli. In consequence, for requirement (*4.a*), we computed a correlation between the indicator values and the corresponding subjective MOS, given that, in this particular case of single degradations, the global subjective MOS reflected entirely the impact of the degradation under consideration. Then, we computed the mapping function for a given quality degradation indicator from its values and the auditory subjective MOS corresponding to the degradation for which it was designed, using a second order polynomial function with a confidence interval of 95%. The coefficients of each quality degradation indicator are available in Table 2. These mapping functions were then used as far as a mapped MOS was required. Concerning requirement (*4.b*), we selected only the worst condition for each dimension (respectively conditions C4, C6, C8 and C10 in Table 1).

The second test allowed evaluating the robustness of the quality degradation indicators for conditions composed of multiple degradations reflecting real communications. For this test, we focused on the behaviour of the mapped MOS value of each quality degradation indicator. In a first step (*Test 2_1*), four conditions (from C11 to C14) were selected in which the degradation relative to the "*DFC*" dimension remained the same for all conditions (use of codec G722.1C with 32kbits/s as bit rate) with a variable Packet Loss (PL) (0%, 2% and 10%) and a speech level attenuation of 10 dB (condition C13). In a second step (*Test 2_2*), three conditions with the same degradation level for "*DFC*" dimension (use of codec G722.2) were considered, including a 10% packet loss and three different noise conditions. These conditions are detailed in Table 1.

Note that, in this study, due to the limited number of conditions, the results for the "f_c", "r_A", "r_I", "*Flatness*", "*Reverb*", and "*FrameRepeat*" indicators are not presented. The study of their relevance will require further speech samples.

6. RESULTS AND DISCUSSION

6.1. Results for single degradation conditions

(1) *DFC quality indicators*: Table 3 shows that the "*Erb*" and "*Freq*" indicators are highly correlated with auditory tests results ($\rho \approx 0.9$) and present a high mapped MOS (≥ 4) for degradations for which they were not dedicated. Based on requirements (*4.a*) and (*4.b*), these quality indicators are relevant.

(2) *Continuity quality indicators*: The "*TimeClip*" and "r_L" indicators perform well in terms of correlation with auditory tests ($\rho \geq 0.9$) (see Table 3) but the "*TimeClip*" indicator presents estimated MOS far below 3.0 (MOS = 1.5) for

Dimensions	Degradation Conditions	Dimensions	Degradation Conditions
	SWB (**C1**)		
DFC	SWB 100-5000 Hz (**C2**)	**DFC**	G.722.1C (32 kbits/s) (**C11**)
	SWB mIRSsend+IRSrcv (**C3**)	**DFC, Continuity**	G.722.1C (32 kbits/s), 2% PL (**C12**)
	SWB 500-2500 Hz (**C4**)	**DFC, Continuity, Loudness**	G.722.1C (32 kbits/s), 2% PL, 10dB$_{SPL}$ (**C13**)
Continuity	SWB 2% time clipping (**C5**)	**DFC, Continuity**	G.722.1C (32kbits/s), 10% PL (**C14**)
	SWB 20% time clipping (**C6**)	**DFC, Continuity, Noisiness**	G722.2, babble (15dB), 10%PL (**C15**)
Noisiness	SWB + 20 dB Babble (**C7**)	**DFC, Continuity, Noisiness**	G722.2, street noise (27dB), 10%PL (**C16**)
	SWB + 12 dB Noise Hoth (**C8**)	**DFC, Continuity, Noisiness**	G722.2, street noise (30dB), 10%PL (**C17**)
Loudness	SWB Level -10 dB (**C9**)		
	SWB Level -20 dB (**C10**)		

Table 1: Summary of conditions used in this study: 10 single degradation conditions (C1,…, C10), 4 conditions for Test2_1 (C11, C12, C13, C14) and 3 conditions for Test2_2 (C15, C16, C17). SWB represents the reference signal. For the condition C3, the reference signal is limited to the band (300-3400 Hz) with IRS (Intermediate Reference System) filtering at send side (mIRSsend) and at received side(IRSrcv)

	α0	α1	α2
Freq	4.19	0.34	-0.091
Erb	-0.768	0.38	-0.0069
TimeClip	5.46	-0.57	0.02
rL	4.32	-77.015	438.53
Noise	4.57	-0.367	0.015
Ln	4.52	-0.081	-0.0003
NoS	4.56	-0.385	0.014
Level	-7682.93	15227.07	-7540.25
Leq	-11.15	0.38	-0.0023
LTL	1.95	0.11	-0.0012

Table 2: Coefficients of the mapping function for each quality degradation indicator obtained from the degradation for which the indicator is designed.

		DFC	Continuity	Noisiness	Loudness
	Indicators				
DFC	*Freq*	0.86	4.42	4.24	4.51
	Erb	0.9	4.41	4.42	4.49
Continuity	*TimeClip*	**1.5**	0.9	4.75	4.75
	rL	4.18	0.9	4.32	4.32
Noisiness	*Noise*	4.077	3.69	0.92	4.48
	Ln	4.5	4.5	0.95	4.51
	NoS	4.32	4.25	0.91	4.41
Loudness	*Level*	4.28	4.48	4.58	**0.7**
	Leq	4.49	4.49	4.49	**0.8**
	Ltl	4.52	4.45	4.42	**0.72**

Table 3: Correlation/Discrimation matrix. The "diagonal" values (colored cells) represent the correlations (ρ) between the quality degradation indicator values and the corresponding auditory MOS (*4.a*). The other values stand for the mapped MOS values of these quality indicators (*4.b*)

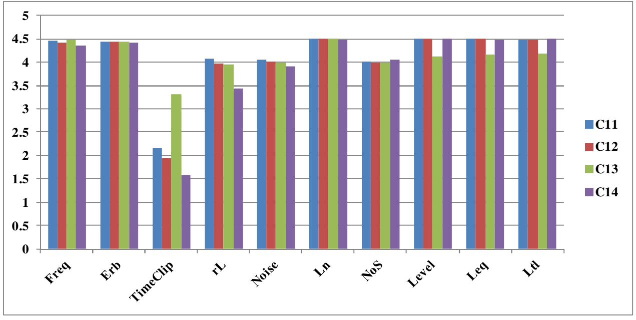

Figure 1: Mapped MOS obtained from Test 2_1

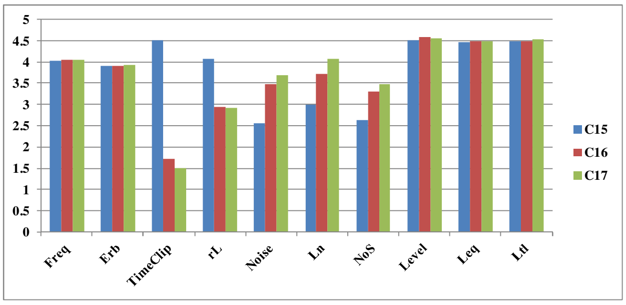

Figure 2: Mapped MOS obtained from Test 2_2

conditions relative to the Directness/Frequency Content dimension. This quality degradation indicator seems to be impacted by the bandwidth limitation and, therefore, is not satisfying to quantify degradations of the Continuity dimension, in contrary to the "r_L" indicator.

(3) *Noisiness quality indicators*: Table 3 shows that the "*Noise*", "*Ln*" and "*NoS*" indicators perform well (ρ ≥ 0.9 and MOS > 3.5) which means that they are relevant for diagnosing degradations linked to the dimension Noisiness.

(4) *Loudness quality indicators*: All quality degradation indicators for this dimension, namely "*Level*", "*Leq*" and "*LTL*", have correlation below 0.9 and thus do not respect the requirement (*4.a*). Nevertheless, the "*Leq*" indicator performs the best (ρ = 0.8) as observed in Table 3 and the mapped MOS values of these three quality degradation indicators are very high (MOS > 4) for degradations for which they were not designed.

6.2. Results for multiple degradations conditions

(1) *DFC quality indicators*: Let us remind that, in the second test (*Test 2_1*), the degradation from the Directness/Frequency Content (DFC) dimension remained the same for all conditions (use of the same codec). Figure 1 shows that the mapped MOS values of the "*Freq*" and "*Erb*" indicators remain relatively the same in spite of the presence of other degradations. The same remark can be made from Figure 2 (*Test 2_2*) which means that these indicators are not impacted by other degradations. Therefore, these quality indicators could be considered as robust to diagnose the

influence of bandwidth limitation, even in the presence of other degradations.

(2) *Continuity quality indicators*: From Figures 1 and 2, it comes out that the "*TimeClip*" indicator displays high variations depending on the tested degradations, which confirms once more that this indicator was not reliable to quantify the effect of discontinuity. As far as the "r_L" indicator is concerned, the mapped MOS values is relatively constant (MOS ≥ 4) for conditions C11, C12 and C13 (*Test 2_1*). However, the performance of this indicator in *Test 2_2* is more difficult to interpret. A difference of 3 MOS points can be observed between condition C15 and conditions C16 and C17.

(3) *Noisiness quality indicators*: The Noisiness dimension was not represented in *Test 2_1*. Figure 1 shows that the "*Noise*", "*Ln*" and "*NoS*" indicators have high mapped MOS (MOS ≥ 4) with respect to requirement (*4.b*). Concerning *Test 2_2,* Figure 2 shows that the evolution of the mapped MOS of these indicators follows the level of background noise and tends to suggest that these quality degradation indicators are less impacted by other degradations.

(4) *Loudness quality indicators*: For conditions where the speech level was not impacted (C11, C12, C14, C15, C16 and C17, see Table 1), the mapped MOS values of the "*Level*", "*Leq*" and "*LTL*" indicators are very high and around 4.0 whereas this MOS is close to 3.7 in condition C13 (see Figure 1). These quality degradation indicators seem robust to quantify the impact of speech level deviation even if they did not respect the requirement (*4.a*) (see Section 6.1).

7. CONCLUSION

In this paper, the performance of several quality degradation indicators in the super-wideband telephony context was analyzed using two technical cause analysis criteria. For conditions showing only one type of degradation, we found that all quality degradation indicators selected in the literature, with one exception ("*TimeClip*"), performed well and should be used for advanced diagnosis of modern telephone networks. Since the performance assessment of quality degradation indicators is a difficult issue when multiple degradations occur simultaneously, we selected then different conditions with mixed degradations, and observed the behaviour of the indicators we selected. We showed that the quality degradation indicators for dimensions DFC, Noisiness and Loudness were somewhat robust to quantify degradations for which they were designed, under the presence of other degradations, which was not the case of indicators theoretically developed to characterize Continuity. This is why, in a future work, we plan to focus on this Continuity dimension and develop and/or optimize indicators to perfectly characterize this dimension.

8. ACKNOWLEGDMENT

We would like to thank the following companies for granting us permission to use their SWB subjectively scored databases and publish results based on them: SwissQual AG, TNO, Deutsche Telekom, Netscout-Psytechnics. The use and analysis of the new P.863 standard model has been made possible thanks to the support of SwissQual and Opticom.

9. REFERENCES

[1] ITU-T Contribution COM12-4, "Speech Degradation Decomposition Using P.862 PESQ Based Approach. 2005-2008", Source: TNO Telecom, Netherlands, International Telecommunication Union, CH Geneva, 2004.

[2] Leman A., Faure J., Parizet E., "Hybrid model for non intrusive speech quality evaluation in telephony applications", 38th International Conference of *Audio Engineering Society conference*, June 2010.

[3] ITU-T P863 Rec. "Perceptual Objective Listening Quality Assessment (POLQA)", *International Telecommunication Union*, CH Geneva, 2011.

[4] Côté N., Koehl V., Gautier-Turbin V., Raake A., Möller S., "An Intrusive Super-Wideband Speech Quality Model: DIAL", In proc. *11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan September 26-30. 2010.

[5] Tiémounou S., Le Bouquin Jeannès R., Barriac V., "Visual Comparison of Perceptual Degradation Indicators in Two Listening Speech Quality Models", In proc. *11th WSEAS International Conference on Signal Processing,* Saint-Malo, France, April 2-4, 2012.

[6] ITU-T Contribution COM12-214, "Benchmark proposal P.TCA", Source: TNO Telecom, Netherlands, International Telecommunication Union, CH Geneva, October 2011.

[7] ITU-T Contribution COM12-53, "POLQA Degradation Decomposition: Perceptual Basis for Degradation Indicators.", Source: Deutsche Telekom AG, International Telecommunication Union, CH Geneva, 2004.

[8] Wältermann M., Scholz K., Raake A., Heute U., Möller S., "Underlying quality dimensions of modern telephone connections", In Proc *9th Int. Conf. on Spoken Language Processing*, pages, 2170-2173, Pittsburgh, USA, September 2006.

[9] McDermott B. J., "Multidimensional Analyses of Circuit Quality Judgments", *Journal of the Acoustical Society of America* 45(3):774-781, 1969.

[10] Scholz K., Wältermann M., Huo, L., Raake A., Möller S., and Heute U., "Estimation of the Quality Dimension "Directness/Frequency Content" for the Instrumental Assessment of Speech Quality", In Proc. *9th International Conference on Spoken Language Processing* (ICSLP), 1523–1526, USA–Pittsburgh, PA, 2006.

[11] Huo L., Wältermann M., Heute U. and Möller S., "Estimation of the Speech Quality Dimension Discontinuity*"*. In Proc. *8th ITG-Fachbericht-Sprachkommunikation*, DE-Aachen, 2008.

[12] Deep S., Lu S., "Objective evaluation of speech signal quality by the prediction of multiple foreground diagnostic acceptability measure attributes". *Journal of Acoustical Society of America.* May 2012.

[13] Huo L., Wältermann M., Heute U. and Möller S., "Estimation Model for the Speech-Quality Dimension Noisiness". *Acoustics'08*, Paris, June 29-July 4, 2008.