Visual Comparison of Perceptual Degradation Indicators in Two Listening Speech Quality Models

SIBIRI TIEMOUNOU$^{1,2}$, REGINE LE BOUQUIN JANNES$^2$, VINCENT BARRIAC$^1$

$^1$Orange Labs/Telecom R&D, Lannion France
$^2$Université de Rennes 1, LTSI, F-35000, France

Campus de Beaulieu, Bât 22, Université de Rennes 1, 35042, Rennes Cedex, France
{sibiri.tiemounou, vincent.barriac}@orange.com, regine.le-bouquin-jeannes@univ-rennes1.fr

Abstract: - Recently, new objective speech quality evaluation methods which integrate perceptual dimensions have been developed in order to provide diagnostic information about specific aspects of the transmission systems' quality. In this paper, we present two approaches, one (DIAL) built explicitly on blocks corresponding to perceptual dimensions and the other one (POLQA) using quality indicators implicitly linked to these perceptual dimensions. The first goal of our study is to identify in POLQA the correspondence between perceptual dimensions and quality indicators. Such indicators can be either present in the model itself or derived from our own analysis of the model. In a second step, we present first results in comparing the indicators of these two approaches. We found that some indicators are equivalent when simple degradations are applied.

Key-Words: - Perceptual dimensions, super wideband, voice quality assessment, objective models

1 Introduction

The well-known objective quality evaluation method PESQ (Perceptual Evaluation of Speech Quality) [1] predicts MOS (Mean Opinion Score) values with a high correlation with auditory tests (also known as subjective tests), but does not provide further diagnostic information about specific aspects of the transmission systems' quality as indicated for instance in [2]. This is why much effort has been done to develop new methods integrating perceptual dimension estimators besides global MOS. Concerning perceptual dimensions, in [2], three types of degradation have been proposed and were proven to be orthogonal: time degradations "Continuity", frequency-response degradations "Directness/Frequency Content" and additive-noise degradations "Noisiness". The existence of a distinct degradation dimension "Loudness" reflecting the impact of speech level on perception is also recognized in [3]. One of these approaches, called DIAL (Diagnostic Intrusive Prediction of Speech Quality) [4], is built explicitly on the 4 perceptual degradation dimensions mentioned above and provides a global predicted MOS-LQO (MOS-Listening Quality Objective) value, together with 4 additional MOS values for each perceptual dimension respectively. Another approach named POLQA (Perceived Objective Listening Quality Assessment) model, now standardized as P.863 [5], implicitly integrates perceptual dimensions that we try firstly to identify in this work. In what follows, we use DIAL as benchmark model against which POLQA is compared. In section 2 of this paper, we identify the degradation dimensions in POLQA. After presenting in section 3 the studied degradation conditions and the corresponding speech material, we analyze in section 4 the relevant indicators in both approaches to quantify and diagnose perceived degradation dimensions.

2 Analysis of POLQA in terms of perceptual dimensions

Although the POLQA model is not based on perceptual dimensions, the presence of several intermediate indicators (mentioned as such in the description provided in P.863) supposes an implicit characterization of these dimensions. In this section, we identify these indicators in POLQA and compare them with our reference model (DIAL) from the point of view of the four perceptual dimensions mentioned above.

2.1 Directness/Frequency Content (DFC)

Also called "Coloration", this dimension is related to the characteristics of the frequency response of the overall transmission system (mouth-to-ear). In DIAL, the DFC dimension is subdivided in 2 sub-
dimensions as proposed in [6], namely "Directness" and "Frequency content". Each sub-dimension is estimated by specific indicators:

(a) Directness: This sub-dimension includes specific impairments such as the influence of talking-room reflections or bandwidth limitation. In DIAL, this sub-dimension is quantified by "ERB" (Equivalent Rectangular Bandwidth). This indicator characterizes the bandwidth limitation of the frequency response. Concerning the POLQA model, the "Reverb" indicator is used to estimate the reverberation effect and is thus rather linked to the "Directness" sub-dimension.

(b) Frequency Content: This sub-dimension is related to the frequency response deviation of overall transmission. In DIAL, the "f.c." indicator, which is the central frequency of the gain of the overall transmission system, is used to quantify this sub-dimension. In POLQA, the "Freq" indicator quantifies the impact of overall global frequency response distortion, similarly to the "FRQ" indicator used in [2]. It addresses both "Directness" and "Frequency content" sub-dimensions. In addition, the "Flatness" indicator quantifies the impact of timbre distortions (also referred as "coloration") and then falls in the scope of this dimension.

2.2 Continuity
This dimension is mostly related to the presence of packet loss and bit errors on the transmission path. In DIAL, this dimension is built on 3 sub-dimensions, partly based on the model developed in [7]:

(a) "Time clipping": It describes the perceived periodicity of the short-time distortions related mainly with IP (Internet Protocol) packet-loss degradations. This sub-dimension is estimated in DIAL by the "r_{l}\) (Short level variation rate) indicator representing the strong level variations in the degraded speech signal mainly caused by IP packet loss. In POLQA, during the calculation of the internal representation of speech signals, a difference between the reference pitch loudness and the degraded pitch loudness reflects apparently the impact of time clipping on the perceived speech quality. We think that this difference could have equivalent meaning with indicators used in DIAL and mostly with the "r_{l}\) indicator. Thus, we derived from this difference an indicator named "TimeClip".

(b) "Additive artifacts": This sub-dimension represents the effect of frame repetition potentially mostly generated by Packet Loss Concealment algorithms. In DIAL, this sub-dimension is quantified by the "r_{a}\) (Artifact rate) indicator. In POLQA, in the same manner as for "TimeClip", during the computation of the final disturbance densities, an indicator is used to estimate severe distortions introduced by frame repetitions, certainly similar, according to us, to the "r_{a}\) indicator used in DIAL. We named it "FrameRepeat".

(c) "Interruptness": This sub-dimension reflects the interruptions introduced in the degraded speech signal when the lost frames are replaced by silence frames and is quantified in DIAL by the indicator "r_{i}\) (Interrupt rate). We found no equivalence of this indicator in POLQA.

2.3 Noisiness
This dimension is related to the perceived impact of noise-like distortions such as background noise or circuit noise.
In DIAL, two indicators are used to quantify this dimension, namely "L_{n}\), which is the total perceived noise loudness computed during silence periods (taking into account abrupt noise level variations), and "NoS" (Noise on Speech), which measures the power of additive noise present during speech active periods. In POLQA, the "Noise" indicator quantifies the impact of additive noise on the whole signal (i.e. during both silent and active speech periods). Besides, let us note that an internal "Noise contrast" indicator, derived from the silent parts of the reference signal, allows quantifying severe noise level variations.

2.4 Loudness
The "Loudness" dimension reflects the degradation for speech heard at a non-optimum listening level. In DIAL, the "LTL" (Long-Term Loudness) indicator estimates the perceived loudness of the whole degraded speech signal. The "Leq" (Equivalent Continuous Sound Level) indicator corresponds to the mean energy of the degraded signal over all active speech frames. In POLQA, the "Level" indicator is used to quantify severe deviations of the optimal listening level. It is applied on the whole signal, thus corresponding more or less to the "LTL" indicator in DIAL.
In this work, we focused on the SWB (Super Wideband) mode of models since our main goal was to identify the root causes of degradations due to VoIP transmission, where NB (Narrow Band), WB (Wide Band), but also more and more SWB codecs are in use. We selected existing databases from the speech data pool of ITU-T study group 12 Question 9 used for the POLQA benchmark, and in particular some new SWB databases specifically developed for the competition. These databases contain 10 common anchor conditions presented in Table 1 (merged per perceptual dimension), each of them (except the reference condition) being impacted on only one of the degradation dimensions. For this study, we worked only with these 10 conditions, yielding a total of 936 speech stimuli taken from 4 databases including 4 languages (French, Dutch, British English and Swiss German).

Note that in this study, the results for the "f<sub>c</sub>"", "r<sub>A</sub>"", "r<sub>I</sub>" indicators from DIAL, "Reverb", "Flatness", "FrameRepeat" and "Noise contrast" indicators from POLQA are not presented here since the corresponding degradations were not present in the selected conditions. The study of their relevance will require further speech samples.

Let us remind that the codes of POLQA and DIAL have been modified so as to provide as outputs (besides MOS scores) all dimension indicators identified above. Concerning POLQA, the consistency in terms of global MOS score of these modifications with a reference implementation (executable used for characterizing the new standard) has been checked.

3 Methods and Material

Table 1: The 10 anchor conditions

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Degradation Conditions</th>
<th>Signals Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFC</td>
<td>SWB 100-5000 Hz (C2)</td>
<td>Reference signal limited to (100-5000 Hz)</td>
</tr>
<tr>
<td></td>
<td>SWB mlIRSsend+IRSrcv (C3)</td>
<td>Reference signal limited to (300-3400 Hz) + IRS (Intermediate Reference System ) filtering, modified at send side</td>
</tr>
<tr>
<td></td>
<td>SWB 500-2500 Hz (C4)</td>
<td>Reference signal limited to (500-2500 Hz)</td>
</tr>
<tr>
<td>Continuity</td>
<td>SWB 2% time clipping (C5)</td>
<td>Reference signal with 2% time clipping</td>
</tr>
<tr>
<td></td>
<td>SWB 20% time clipping (C6)</td>
<td>Reference signal with 20% time clipping</td>
</tr>
<tr>
<td>Noisiness</td>
<td>SWB + 20 dB Babble (C7)</td>
<td>Reference signal with non stationary babble noise</td>
</tr>
<tr>
<td></td>
<td>SWB + 12 dB Noise Hoth (C8)</td>
<td>Reference signal with stationary Hoth noise</td>
</tr>
<tr>
<td>Loudness</td>
<td>SWB Level -10 dB (C9)</td>
<td>Reference signal attenuated by 10 dB</td>
</tr>
<tr>
<td></td>
<td>SWB Level -20 dB (C10)</td>
<td>Reference signal attenuated by 20 dB</td>
</tr>
</tbody>
</table>
4 Experimental results

In order to obtain reliable results, speech stimuli have been grouped by condition in Figures 1 to 4. This involves that all conditions are represented by the same number of male speech samples, female speech samples and languages.

4.1 Directness/ Frequency Content (DFC)

The conditions used for this dimension are representative of a simple bandwidth limitation (conditions C2 and C4, when the reference is condition 1), with an exception (condition C3), where the signal is also IRS-filtered (Intermediate Reference System-filtered), even if its bandwidth is between those of conditions C2 and C4. In Figure 1.a, we can observe that the "Freq" indicator increases according to the degree of bandwidth restriction. As for the "ERB" indicator (see Figure 1.b), it depends on the bandwidth of the frequency response of the overall system and then decreases according to the degradation degree. So, from these observations, we can conclude that the "Freq" and "ERB" indicators could be relevant for assessing the degradation linked to bandwidth limitation.

4.2 Continuity

The conditions C5 and C6 used for this dimension represent two steps of simple (i.e. randomly distributed) time clipping. We can observe in Figure 2 that the "rL" indicator of DIAL which quantifies the impact of packet loss is obviously relevant for the dimension "Continuity". The same holds for the "TimeClip" estimator in POLQA (see Figure 2). The respective differences between values reached by these indicators are clearly reflecting the respective differences between loss percentages.

4.3 Noisiness

The major difference with this dimension, compared to the other dimensions, is that there is no real continuum between the two studied conditions (C7 and C8) in terms of magnitude of a given type of degradation. They differ from each other not only by their level of noise (12 and 20 dB of Signal on Noise ratio), but also by the type of noise (one with steady type, the other one with a time-varying type). This makes the analysis of the corresponding indicators more difficult. However, Figure 3 shows that all the selected indicators vary according to the level of additive noise. The "Noise" and "Lm" indicators are the most stable ones for a given condition. Since the POLQA "Noise" indicator is estimated on the whole degraded signal, we ask ourselves whether a combination of the two estimators derived from DIAL, i.e. "NoS" (estimated on active speech frames) and "Lm" (estimated on silence periods), could reflect the behavior of this "Noise" indicator.

4.4 Loudness

The conditions used for the "Loudness" dimension (C9 and C10) represent two steps (of 10 dB each) of signal attenuation compared to the original speech sample, applied uniformly on the whole signal. Note that the "Level" indicator is computed as a ratio of an optimum level power density by the power density of the degraded signal. This involves, of course, that it increases when the level power density of the degraded signal decreases, or, in other words, when its level is attenuated (see Figure 4.a). Concerning the indicators from DIAL, namely "LTL" and "Leq", the first one is computed on the whole degraded signal and the second one only on the active speech frames. Consequently, they decrease when the level of degraded signal decreases (see Figure 4.b). Since the attenuation applied in these conditions is the same for all parts of signal, it is logical that "LTL" and "Leq" convey comparable information. To study their differences, signals with different levels of attenuation on active and silent parts of the signal should be used.

5 Conclusion and Perspectives

In this paper, psycho-acoustic models (POLQA and DIAL) have been studied in order to find how they could address perceptual degradation dimensions and in particular which implicit or explicit indicators could be used for diagnostic on the causes for these degradations. To do so, we used different databases of speech stimuli containing the same anchor degradation conditions which are representative of the four main perceptual dimensions, namely Directness/Frequency Content, Continuity, Noisiness and Loudness. As a result of our study, we found that some indicators from both models are relevant and could be used to diagnose the perceived degradations. Most implicit indicators in POLQA are equivalent to explicit ones in DIAL. Since the degradation conditions we studied were limited, these conclusions need an investigation on a larger number of speech samples and databases. This will allow us to test the reliability and the robustness of the studied indicators and complete the present study with the whole set of indicators we identified. In a future step, we first plan to focus on
the "Continuity" and the "Directness/Frequency Content" dimensions. As a matter of fact, the "Noisiness" dimension has been largely investigated in the past years [8], whereas the "Loudness" dimension seems easier to characterize and does not need any priority investigation.

6 Acknowledgement
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 POLQA standard model has been made possible thanks to the support of SwissQual and Opticom.

References: