Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 1047–1063

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

On the identification of relevant degradation indicators in super wideband listening quality assessment models

Sibiri Tiémounou [a,b,c,*], Régine Le Bouquin Jeannès [b,c], Vincent Barriac [a]

[a] Orange Labs–Lannion, 2 Av. Pierre Marzin, 22307 Lannion Cedex, France
[b] INSERM, U 1099, Rennes F-35000, France
[c] Université de Rennes 1, LTSI, Rennes F-35000, France

Received 23 January 2013; received in revised form 10 June 2013; accepted 13 June 2013
Available online 2 July 2013

Abstract

Recently, new objective speech quality evaluation methods, designed and adapted to new high voice quality contexts, have been developed. One interest of these methods is that they integrate voice quality perceptual dimensions reflecting the effects of frequency–response distortions, discontinuities, noise and/or speech level deviations respectively. This makes it possible to use these methods also to provide diagnostic information about specific aspects of the transmission systems' quality, as perceived by end-users. In this paper, we present and analyze in depth two of these approaches namely POLQA (Perceived Objective Listening Quality Assessment) and DIAL (Diagnostic Instrumental Assessment of Listening quality), in terms of quality degradation indicators related to the perceptual dimensions these models could embed. The main goal of our work is to find and propose the most robust quality degradation indicators to reliably characterize the impact of degradations relative to the perceptual dimensions described above and to identify the underlying technical causes in super wideband telephone communications [50, 14000] Hz. To do so, the first step of our study was to identify in both models the correspondence between perceptual dimensions and quality degradation indicators. Such indicators could be either present in the model itself or derived from our own investigation of the model. In a second step, we analyzed the performance and robustness of the identified quality degradation indicators on speech samples only impaired by one degradation (representative of one perceptual dimension) at a time. This study highlighted the reliability of some of the quality degradation indicators embedded in the two models under study and stood for a first step in the evaluation of performance of these indicators to quantify the degradation for which they were designed.

Keywords: Perceptual dimensions; Super wideband; Voice quality assessment; Objective models; Diagnostic

1. Introduction

With the introduction of transport of information on packet networks, the old bandwidth limitations of TDM (Time Division Multiplexing) networks have vanished. In particular, the coding of audio signals is no longer restricted to PCM (Pulse Code Modulation) frames coded at 8 kHz. Thus, new audio contents, without theoretical constraint on their temporal and spectral characteristics, can be carried onto telecommunication networks, opening the door to new services and new customer experience. Amongst these services, high quality telephony is promised to a fruitful future. Nowadays, Wideband (WB) (*i.e.* [50–7000] Hz) telephony, also known as HD (High Definition) Voice (Danno, 2013), is more and more common, and there are even some soft-telephony solutions allowing a so-called super wideband (SWB) telephony, corresponding to frequencies up to 14 kHz (*i.e.* [50–14000] Hz) and a sampling rate at 32 kHz or beyond.

The user experience of such services completely differs from what we have been facing until now. For instance, rich musical contents, or unfiltered environmental noise,

* Corresponding author at: Orange Labs–Lannion, 2 Av. Pierre Marzin, 22307 Lannion Cedex, France. Tel.: +33 296053645.
E-mail addresses: sibiri.tiemounou@orange.com (S. Tiémounou), regine.le-bouquin-jeannes@univ-rennes1.fr (R.L.B. Jeannès), vincent.barriac@orange.com (V. Barriac).

0167-6393/$ - see front matter
http://dx.doi.org/10.1016/j.specom.2013.06.010

can be carried together with or without voice. Some audible defects in the audio signal, particularly present in the higher frequencies, become annoying in SWB context whereas they are attenuated and hardly perceptible in Narrowband (NB) context (*i.e.* [300–3400] Hz). This explains why the current knowledge on quality expectations and perception has to be revisited.

Besides, the complexity of transmission and signal processing scenarios makes it harder and harder to troubleshoot end-to-end quality problems. New diagnosis tools are therefore needed by operational teams to help them in targeting the sources of degradations.

To achieve these two goals, the development of new tests and measurement methods (or the update of existing ones) is required. As far as voice quality measurement is concerned, the well-known PESQ (Perceptual Evaluation of Speech Quality, ITU-T Recommendation P.862 (ITU-T Rec, 2001) listening quality model, for instance, predicts MOS (Mean Opinion Score) values with a high correlation with auditory tests results (also known as subjective tests (ITU-T Rec, 1996) in Narrowband audio telephony context, but such a method presents several drawbacks. It is not able to cope with SWB telephony (and its Wideband extension called WB-PESQ (ITU-T Rec, 2005) presents some limitations (Beerends and Van Vugt, 2004; Côté et al., 2006), and it does not provide further diagnostic information about specific aspects of the transmission systems' quality, as indicated for instance in Beerends and Van Vugt (2004). To overcome these obstacles, ITU-T study Group 12 decided to develop a new method (project named POLQA), adapted to the new audio contents (*e.g.* including super wideband audio contents) and potentially integrating, beside global quality scores, estimators of speech quality features grouped in what we call hereafter "perceptual dimensions" (because of their similar impact on speech quality as perceived by human ear).

With regard to voice quality perception, a distinction between three types of perceptual dimension, associated with corresponding degradations (proven to be orthogonal) has been already proposed in Wältermann et al. (2006). The first dimension, called "Directness/Frequency Content" (Scholz et al., 2006), or "Coloration" (Wältermann et al., 2008), is related to frequency–response degradations. Results obtained by Leman et al. (2010) in Narrowband telephony context revealed that the "Coloration" dimension is mostly linked to coding. Consequently, with a view to diagnosis and to deal with all possible degradations, coding was included in this dimension.

The "Continuity" dimension corresponds to degradations in the time domain, such as loss of signal, clipping, and sudden level variations. Finally, "Noisiness" is linked to perceived noise degradations. The existence of a fourth perceptual dimension reflecting the impact of the speech level, (called "Loudness") has been recognized in McDermott (1969). These four perceptual dimensions are assumed to cover the whole speech quality space including modern telecommunication networks and services.

To estimate quality degradation on these dimensions in the new telephony context and troubleshoot problems, Quality Degradation Indicators (QDIs) must be developed. These indicators are designed to quantify the underlying technical causes for a specific degradation perceived on speech quality. Instead of building them from scratch, we found it wiser to look first at those already developed, and then built new ones only if the study of existing ones concluded that they could not be used as reliable and robust diagnostic tools. When looking at the state-of-the-art, it became quickly obvious that the best place to dig out such indicators was inside signal-based voice quality models. When we started our study, ITU-T was just ending the POLQA (Perceived Objective Listening Quality Assessment) project launched in 2007, aiming at standardizing a new objective perceptual speech quality evaluation model to overcome PESQ, with a broader scope (in particular SWB speech). One of the 6 candidate models called DIAL (Diagnostic Instrumental Assessment of Listening quality) (Côté, 2011) has been developed jointly by FT (France Telecom) and DT (Deutsche Telekom). DIAL was explicitly built on the 4 perceptual degradation dimensions mentioned above and provided a global predicted MOS-LQOsw (MOS-Listening Quality Objective in super wideband context) value, together with 4 additional MOS values for each perceptual dimension respectively.

Once the essential parts of the code of the winner of the POLQA competition (now standardized as ITU-T P.863 Recommendation (ITU-T Rec, 2011), and simply called POLQA in what follows), had been made public, we included it also in our study. Note that, since there is not a directly compilable version of POLQA available, we had to implement the missing blocks and build an executable version. In this latter case, the perceptual dimensions were implicitly integrated in the model.

The goal of our study was to find and propose QDIs in order to quantify the impact of the degradations on the four dimensions exposed above as well as to identify the underlying technical causes in super wideband context. Such indicators are potentially, but not necessarily, extracted from the two models introduced above. However, in this paper, we only present and analyze indicators found in the codes of DIAL and POLQA. It is important to underline that we considered these two objective models in our study for two reasons: (1) they stand for the first objective models to characterize the voice quality impairments in super wideband context whose codes were available and (2) they explicitly or implicitly integrate the four perceptual dimensions described above. Other objective models explicitly based on the 3 perceptual dimensions ("Directness/Frequency Content", "Continuity" and "Noisiness") were proposed in Wältermann et al. (2008), Leman et al. (2010) but these models were limited to Narrowband context. Note that our work did not intend to compare these two models from the point of view of their performance to predict the speech quality, but was rather to select the most relevant QDIs that are robust enough