OPEN ACCESS Freely available online

Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II–Perceptual Model

JOHN G. BEERENDS,[1] *AES Fellow,* **CHRISTIAN SCHMIDMER[2], JENS BERGER[3], MATTHIAS OBERMANN[2], RAPHAEL ULLMANN[3], JOACHIM POMY[2], AND MICHAEL KEYHL,[2]** *AES Member*

[1]*TNO, P. O. Box 5050, NL-2600 GB Delft, The Netherlands*
[2]*OPTICOM GmbH, Nägelsbachstrasse 38, D - 91052 Erlangen, Germany*
[3]*SwissQual AG, Allmendweg 8, CH-4528 Zuchwil, Switzerland*

In two closely related papers we present POLQA (Perceptual Objective Listening Quality Assessment), the third generation perceptual objective speech quality measurement algorithm, standardized by the International Telecommunication Union (ITU-T) as Recommendation P.863 in 2011. This measurement algorithm simulates subjects that rate the quality of a speech fragment in a listening test using a five-point opinion scale. The new standard provides a significantly improved performance in predicting the subjective speech quality in terms of Mean Opinion Scores when compared to PESQ (Perceptual Evaluation of Speech Quality), the second generation of objective speech quality measurements. The new POLQA algorithm allows for predicting speech quality over a wide range of distortions, from "High Definition" super-wideband speech (HD Voice, audio bandwidth up to 14 kHz) to extremely distorted narrowband telephony speech (audio bandwidth down to 2 kHz), using sample rates between 48 and 8 kHz. POLQA is suited for distortions that are outside the scope of PESQ such as linear frequency response distortions, time stretching/compression as found in Voice-over-IP, certain types of codec distortions, reverberations, and the impact of playback volume. POLQA outperforms PESQ in assessing any kind of degradation making it an ideal tool for all speech quality measurements in today's and future mobile and IP based networks. This paper (Part II) outlines the core elements of the underlying perceptual model and presents the final results.

0 INTRODUCTION

During the past decades objective speech quality measurement methods have been developed and deployed using a perceptual measurement approach. In this approach a perception-based algorithm simulates the behavior of a subject that rates the quality of an audio fragment in a listening test. For speech quality one mostly uses the so-called absolute category rating listening test, where subjects judge the quality of a degraded speech fragment without having access to the clean reference speech fragment. Listening tests carried out within the International Telecommunication Union (ITU) mostly use an absolute category rating (ACR) five-point opinion scale [1], [2] that is consequently also used in the objective speech quality measurement methods that were standardized by the ITU, PSQM (Perceptual Speech Quality Measure, ITU-T Rec. P.861, 1996) [3], [4], and its follow-up PESQ (Perceptual Evaluation of Speech Quality, ITU-T Rec. P.862, 2000) [5] – [9]. The focus of these measurement standards is on narrowband speech quality (audio bandwidth 100–3500 Hz) [3] – [8], although a wideband extension (50–7000 Hz) was devised in 2005 [9]. PESQ provides for very good correlations with subjective listening tests on narrowband speech data and acceptable correlations for wideband data.

As new wideband voice services are being rolled out by the telecommunication industry, the need emerged for an advanced measurement standard of verified performance and capable of higher audio bandwidths [10]. Therefore ITU-T (ITU-Telecom sector) Study Group 12 initiated the standardization of a new speech quality assessment algorithm as a technology update of PESQ. The new, third

generation measurement standard, POLQA (Perceptual Objective Listening Quality Assessment), overcomes shortcomings of the PESQ P.862 standard such as incorrect assessment of the impact of linear frequency response distortions, time stretching/compression as found in Voice-over-IP, certain types of codec distortions, and reverberations. Furthermore, POLQA allows assessing the impact of playback level and can deal with super-wideband speech (14 kHz audio bandwidth). POLQA was accepted in January 2011 by ITU-T as Rec. P.863 [11].

This paper (Part II) provides an overview of the subjective testing procedure (Section 1) and the perceptual model (Section 2) used in the POLQA standard, including the performance of the new standard (Sections 3 and 4) and the most important conclusions (Section 5). The temporal alignment, including the model requirements and basic modeling approach, are given in Part I.

1 SUBJECTIVE SPEECH QUALITY EXPERIMENTS

The development of POLQA required large amounts of reliable subjective data for narrow, wide and super wideband speech signals. A complete description of the construction of the databases and the subjective test procedure is beyond the scope of this paper, but the main points are given in the next paragraphs. A detailed overview of the subjective test procedure is given in Appendix II of ITU-T Recommendation P.863 [11].

In order to be able to assess high quality voice systems, the reference recordings that are used in both the subjective and objective testing should be of the highest quality. Therefore strict requirements were set on the voice recordings that are used as the reference files. This reference material must be recorded in a low reverberant room (reverberation time below 300 ms above 200 Hz), preferably an anechoic room. Recordings are made using an omnidirectional microphone with a distance between mouth and microphone of approximately 10 cm. The A-weighted sound pressure level (A-SPL) re 20 μPa of the background noise should be below 30 dB, and the A-weighted RMS power relative to the digital overload point of the noise floor of the final recordings should not exceed –84 dB. Speech signals are sampled at 48 kHz and band pass filtered between 50 Hz and 14 kHz.

Each reference speech file consists of two sentences separated by a gap of at least 1 s but not more than 2 s. The minimum amount of active speech in each file is 3 s. The first speech activity starts between 0.5 and 2 s. The last speech activity ends between 0.5 and 2.5 s before the end of the speech file. A minimum set of 16 different reference samples are required and no repetition of texts is allowed in this set and the samples use at least four different speakers. The digital Active Speech Level (ASL according to ITU-T P.56 [12]) of the signals has to be −26 dBov (dB overload) for presentation at the nominal level. The corresponding nominal Sound Pressure Level (SPL) in the acoustical domain is 73 dB at the Ear Reference Point (ERP) using a diffuse field equalized diotic headset (same signal to both ears) for super-wideband or 79 dB at the ERP using an Intermediate Reference System (IRS) type telephone handset (monotic presentation). For all tests in super-wideband mode, level variations between −20 to +6 dB relative to the nominal level were used, allowing to investigate the impact of playback level on perceived speech quality.

Table 1. ACR listening quality opinion scale [1], [2] used in the development of POLQA. The average score over a large set of subjects is called MOS-LQS (Mean Opinion Score Listening Quality Subjective).

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

The listening tests use an absolute category rating (ACR) type of test where subjects listen to speech files and express their opinion on a rating scale. Within the telecom industry one mostly uses a five-point opinion scale [1], [2] (see Table 1). Up to now most subjective tests used narrowband speech (maximum audio bandwidth 100–3500 Hz) as the best possible quality, resulting in an overestimation of the quality of the degraded speech. In wideband testing the best quality speech has an audio bandwidth of 50–7000 Hz, while in super-wideband tests this is extended toward 50–14000 Hz. One should be aware of the fact that subjects tend to adapt their opinion rating toward the maximum quality used in the subjective test. This results in the effect that a high quality narrowband speech file in a narrowband test will get a higher MOS score in comparison to when this file is presented in a (super-)wideband experiment. About 24 naive subjects were used in each test, between 15 and 60 years old, equally distributed over the categories 15–30, 31–50, 51–65, and male/female. The final results of a subjective test are expressed in terms of Mean Opinion Scores for Listening Quality Subjective (MOS-LQS).

In order to be able to compare the results of different subjective tests, the following 12 reference anchor conditions are included in each of the subjective tests:

- Clean 0 dB, −10 dB, and –20 dB relative attenuation;
- Multiplicative noise (MNRU conditions [13]) with signal-to-noise ratios of 10 dB and 25 dB (using P.50 [14] shaped noise for modulation);
- Additive noise with a signal-to-noise ratio of 12 dB using Hoth noise [15] and 20 dB using babble noise;
- Linear filtering with narrowband telephone characteristic (300–3400 Hz), bandpass filters 500–2500 Hz, and 100–5000 Hz;
- Temporal clipping with 2% and 20% packet loss, packet size 20 ms without packet loss concealment.

The ITU-T Study Group 12 POLQA benchmark consisted of three phases—a model training phase, a model

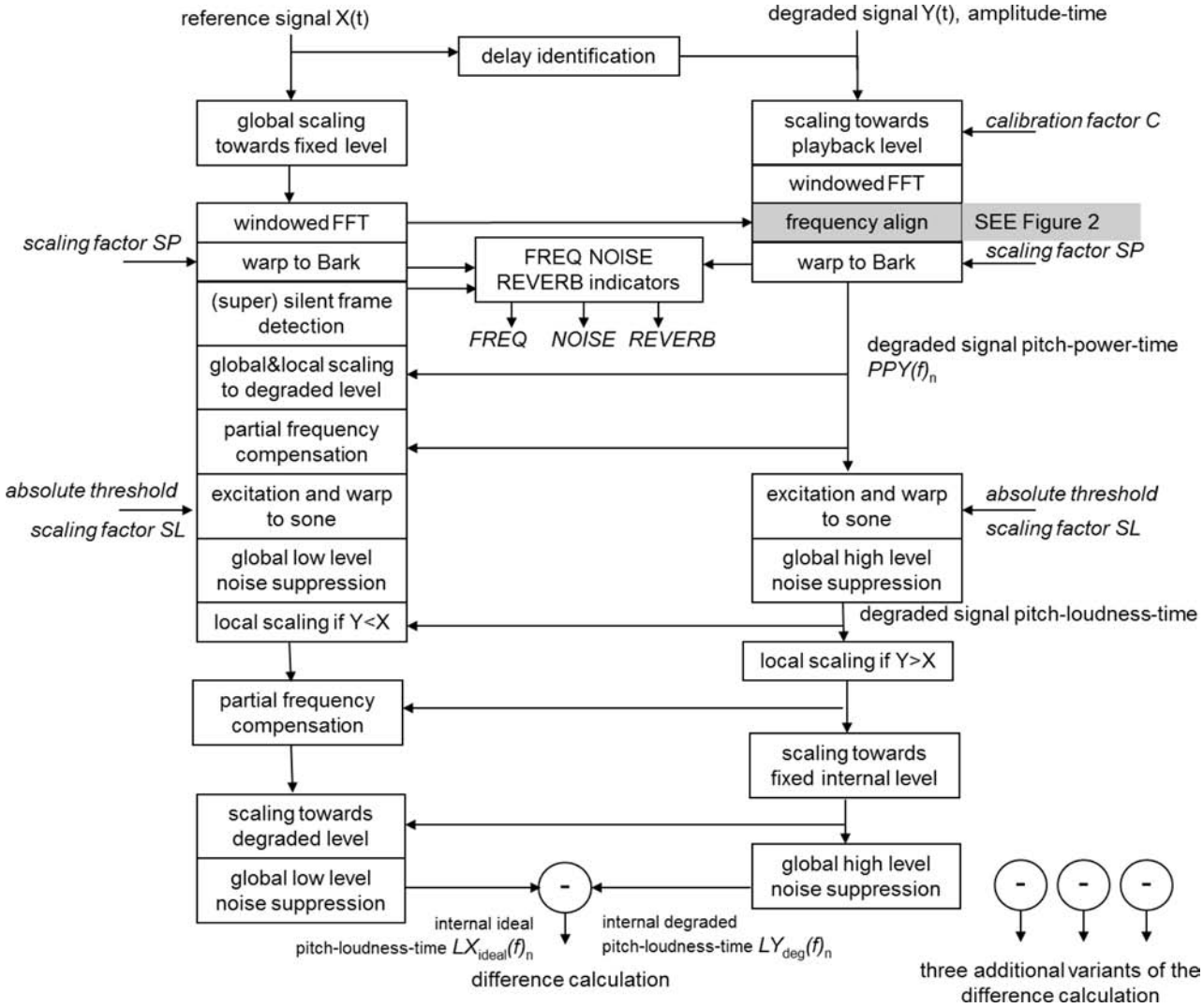

Fig. 1. Overview of the first part of the POLQA perceptual model: Calculation of the internal representation of the reference and degraded signals (see Sections 5.1 through 5.10). Four different variants of the internal representations are calculated (represented by the four circles with a – sign), each focused on a specific set of distortions (see Sections 2.11 and 2.12).

validation and selection phase, and finally a model integration phase. In total 16 wideband/super-wideband databases were used in the training phase of POLQA while the objective model validation and selection was carried out with 8 wideband/super-wideband databases. Additional information on the super-wideband databases is given in [16]. Besides these tests, POLQA was also trained and validated with respectively 29 and 9 narrowband databases for checking the performance in narrowband mode with standard telephone handset playback. This allows comparison of the performance of POLQA P.863 with PESQ P.862 in the classical narrowband (300–3400 Hz) telephone listening situation. In the training phase all training data were made available to all proponents while the databases used in the model validation and selection were not available to any of the proponents and for the majority created after the models were submitted to the ITU-T. The model selection was thus carried out on data which none of the submitted models had "seen."

2 POLQA PERCEPTUAL MODEL

As explained in Part I Section 3, the basic approach of POLQA (ITU-T rec. P.863) is the same as used in PESQ (ITU-T Rec. P.862), i.e., a reference input and degraded output speech signal are mapped onto an internal representation using a model of human perception. The difference between the two internal representations is used by a cognitive model to predict the perceived speech quality of the degraded signal. An important new idea implemented in POLQA is the idealization approach that removes low levels of noise in the reference input signal and optimizes the timbre. Further major changes in the perceptual model include the modeling of the impact of playback level on the perceived quality and a major split in the processing of low and high levels of distortion.

An overview of the perceptual model used in POLQA is given in Figs. 1 through 4. Fig. 1 provides the first part of the perceptual model used in the calculation of the

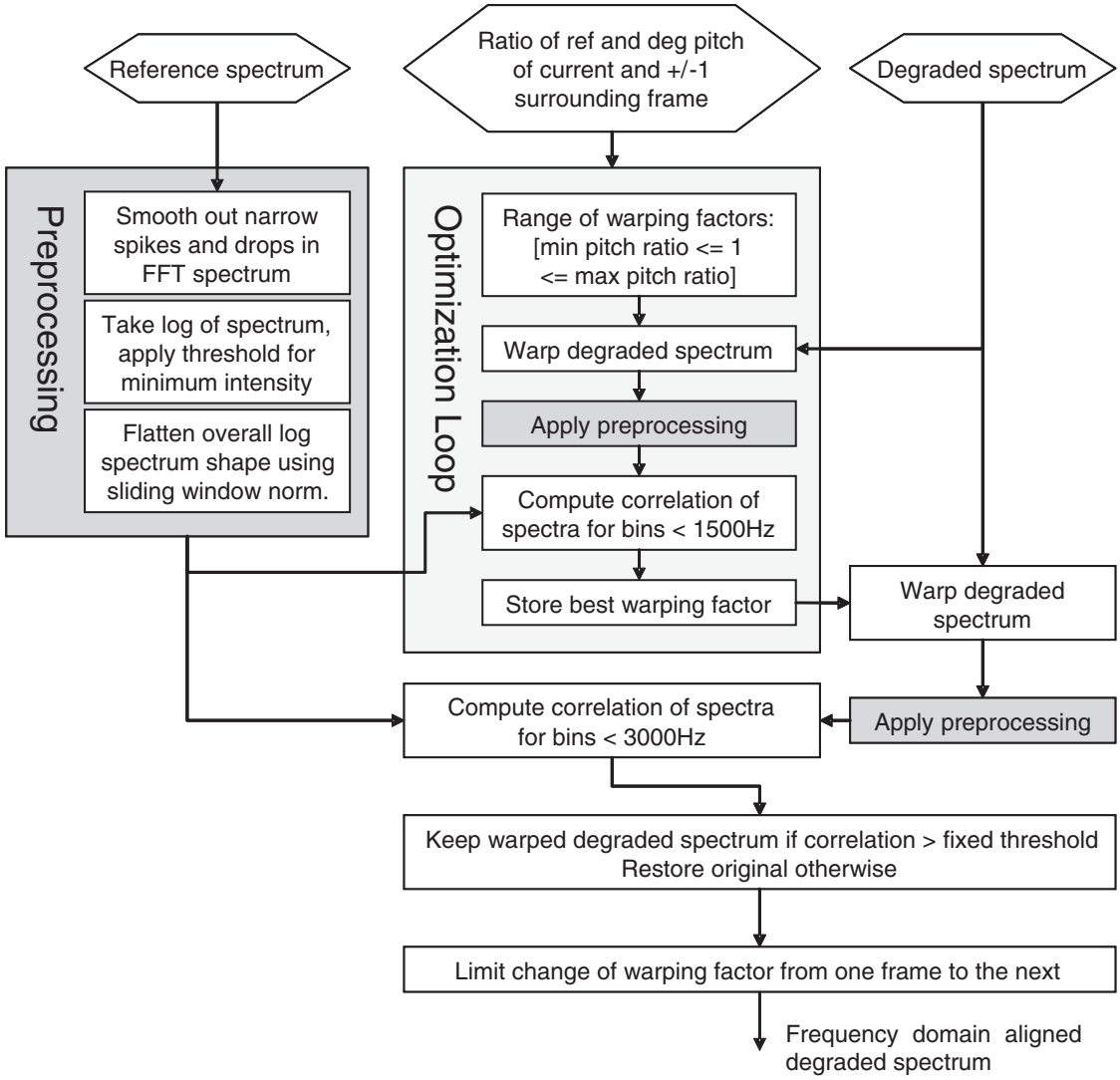

Fig. 2. Overview of the frequency domain alignment used in the POLQA perceptual model.

internal representation of the reference input signal X(t) and the degraded output signal Y(t). Both are scaled and the internal representations in terms of pitch-loudness-time are calculated in a number of steps described in Sections 2.2 through 2.11 after which a difference function is calculated, indicated in Fig. 1 with the left-most circle with a – sign in the center. Like in PESQ [8] two different flavors of the perceptual difference function are calculated, one for the overall disturbance introduced by the system under test and one for the added parts of the disturbance. This models the asymmetry in impact between degradations caused by leaving out time-frequency components from the reference signal and degradations caused by the introduction of new time-frequency components [17]. In POLQA both flavors are calculated in two different approaches, one focused on the normal range of degradations and one focused on loud degradations resulting in four difference function calculations indicated in Fig. 1 with the four circles with a – sign in the center.

For degraded output signals with frequency domain warping, an alignment algorithm is used given in Fig. 2. The final processing for getting the MOS-LQO scores is given in Figs. 3 and 4.

POLQA starts with the calculation of some basic constant settings (Section 2.1) after which the pitch power densities (power as function of time and frequency) of reference and degraded signals are derived from the time- and frequency-aligned time signals (Section 2.2). From the pitch power densities the internal representations of reference and degraded signals are derived in a number of steps as described in Sections 2.3 through 2.11. Furthermore, these densities are also used to derive the first three POLQA quality indicators (Section 2.4) for frequency response distortions (FREQ), additive noise (NOISE), and room reverberations (REVERB). These three quality indicators are calculated separately from the main disturbance indicator in order to allow a balanced impact analysis over a large range of different distortion types. These indicators can also be used

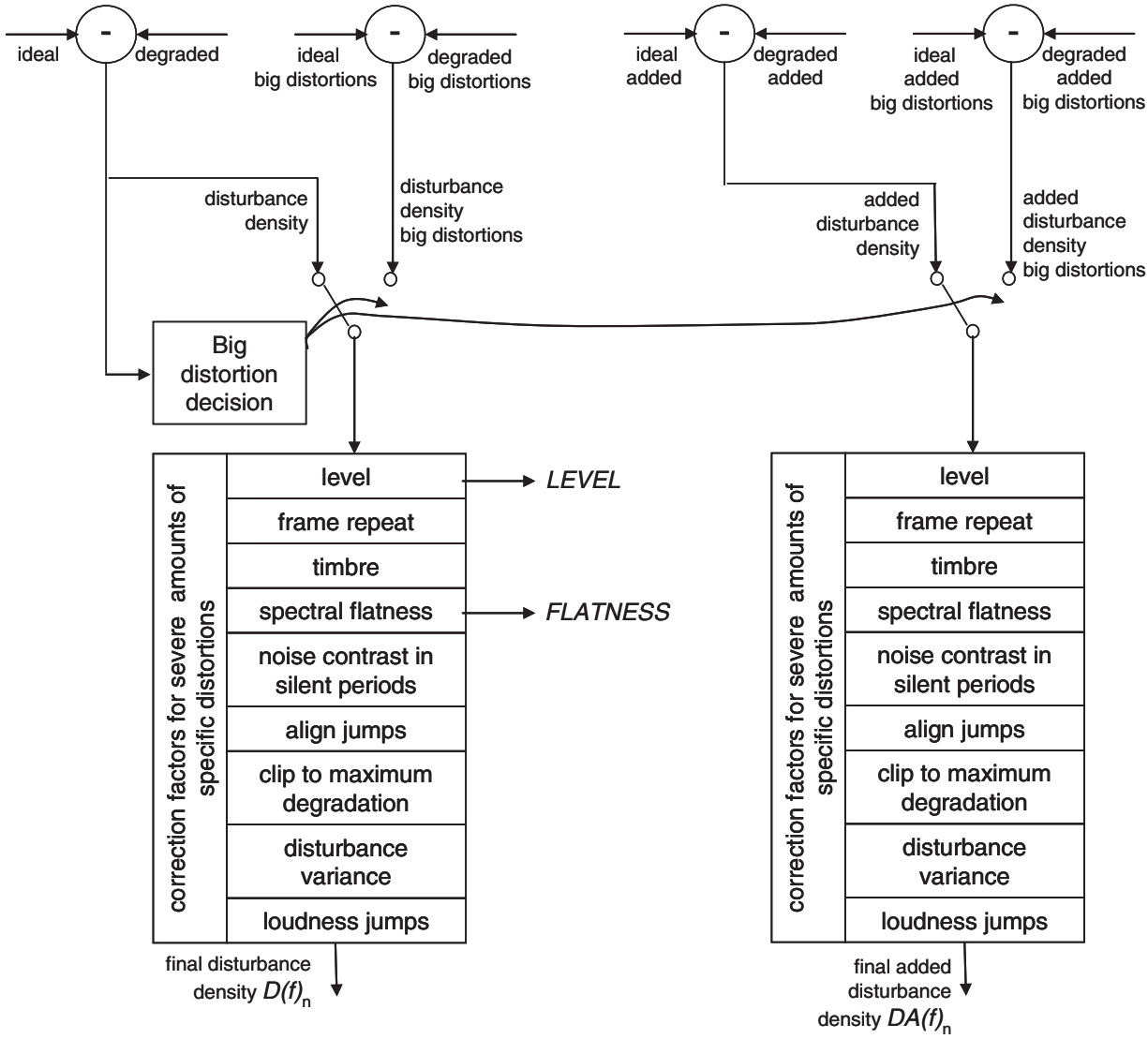

Fig. 3. Overview of the second part of the POLQA perceptual model. Calculation of the final disturbance densities from the four different variants of the internal representations distortions (see Sections 2.11 and 2.12).

for a more detailed analysis of the type of degradations that were found in the speech signal using a degradation decomposition approach as given in [18].

As stated before, four different variants of the internal representations of reference and degraded are calculated, two variants focused on the disturbances for normal and big distortions, and two focused on the added disturbances for normal and big distortions. These four different variants (the circles with a – sign in the center) are the inputs to the calculation of the final disturbance densities as given in Fig. 3.

The internal representations of the reference are referred to as ideal representations because low levels of noise in the reference are removed and timbre distortions as found in the degraded signal that resulted from a non-optimal timbre of the original reference recordings are taken into account. The deviation from the optimal timbre is quantified by a loudness difference between a lower and an upper Bark band of the degraded signal and "punishes" any severe imbalance irrespective of the fact that this could be the result of an incorrect voice timbre of the reference speech file (see Section 2.12). Note that a transparent chain using poorly recorded reference signals, containing too much noise and/or an incorrect voice timbre, will thus not provide the maximum MOS score in a POLQA end-to-end speech quality measurement.

The four different variants of the ideal and degraded internal representations are used to calculate two final disturbance densities, one representing the final disturbance as a function of time and frequency focused on the overall degradation, and one representing the final disturbance as a function of time and frequency but focused on the processing of added degradations (Section 2.12).

Fig. 4 gives an overview of the calculation of the MOS-LQO, the objective MOS score, from the two final disturbance densities, and the FREQ, NOISE, REVERB indicators (Sections 2.13 and 2.14).

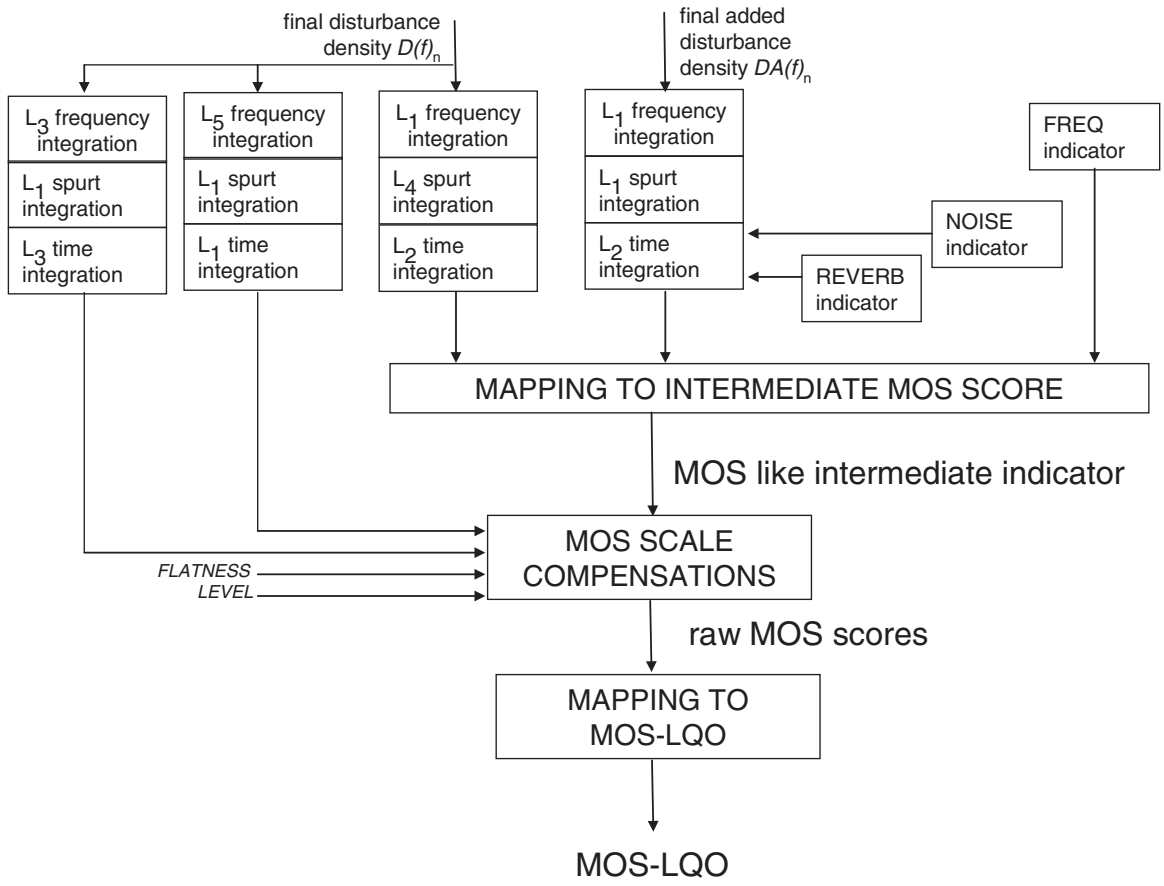

Fig. 4. Overview of the third part of the POLQA perceptual model. Calculation of the final objective listening quality MOS score (MOS-LQO) from the final disturbance densities (see Sections 2.13 and 2.14).

2.1 Pre-Computation of Constant Settings

2.1.1 FFT Window Size Depending on the Sample Frequency

POLQA operates on three different sample rates, 8, 16, and 48 kHz sampling for which the window size W is set to respectively 256, 512, and 2048 samples in order to match the time analysis window of the human auditory system [19]. The overlap between successive frames is 50% using a Hann window. The power spectra—the sum of the squared real and squared imaginary parts of the complex FFT components—are stored in separate real valued arrays for both the reference and the degraded signal. Phase information within a single frame is discarded in POLQA and all calculations are based on the power representations only.

2.1.2 Start Stop Point Calculation

In subjective tests, noise will usually start before the beginning of the speech activity in the reference signal. However, one can expect that leading steady state noise in a subjective test decreases the impact of steady state noise, while in objective measurements that take into account leading noise it will increase the impact; therefore it is expected that omission of leading and trailing noises is the correct perceptual approach. Therefore, after having verified the expectation in the available training data, the start and stop points used in the POLQA processing are calculated from the beginning and end of the reference file. The sum of five successive absolute sample values (using the normal 16 bits PCM range $\pm$32,768) must exceed 500 from the beginning and end of the original speech file in order for that position to be designated as the start or end. The interval between this start and end is defined as the active processing interval. Distortions outside this interval are ignored in the POLQA processing.

2.1.3 The Power and Loudness Scaling Factor SP and SL

For calibration of the FFT time-to-frequency transformation, a sine wave with a frequency of 1000 Hz and an amplitude of 40 dB SPL is generated, using a reference signal X(t) calibration toward 73 dB SPL. This sine wave is transformed to the frequency domain using a windowed FFT with a length determined by the sampling frequency. After converting the frequency axis to the Bark scale the peak amplitude of the resulting pitch power density is then normalized to a power value of 10^4 by multiplication with a power scaling factor SP.

The same 40 dB SPL reference tone is used to calibrate the psychoacoustic (Sone) loudness scale. After warping the intensity axis to a loudness scale using Zwicker's law [20] the integral of the loudness density over the Bark frequency

scale is normalized to 1 Sone using the loudness scaling factor *SL*.

2.2 Scaling and Calculation of the Pitch Power Densities

The degraded signal Y(t) is multiplied by the calibration factor *C* that takes care of the mapping from dB overload in the digital domain to dB SPL in the acoustic domain and then transformed to the time-frequency domain with 50% overlapping FFT frames. The reference signal X(t) is scaled toward a predefined fixed optimal level of about 73 dB SPL equivalent before being transformed to the time-frequency domain. This calibration procedure is fundamentally different from the one used in PESQ [8] where both the degraded and reference are scaled toward a predefined fixed optimal level. PESQ pre-supposes that all play out is carried out at the same optimal playback level while in the POLQA subjective tests levels between −20 dB to +6 to relative to the optimal level are used. In the POLQA perceptual model one can thus not use a scaling toward a predefined fixed optimal level.

After the level scaling, the reference and degraded signal are transformed to the time-frequency domain using the windowed FFT approach described in Section 2.1.1. For files where the frequency axis of the degraded signal is warped when compared to the reference signal, a dewarping in the frequency domain is carried out on the FFT frames as shown in Fig. 2. In the first step of this dewarping, both the reference and degraded FFT power spectra are preprocessed to reduce the influence of both very narrow frequency response distortions, as well as overall spectral shape differences on the following calculations. The preprocessing consists in smoothing, compressing, and flattening the power spectrum. The smoothing operation is performed using a sliding window average of the powers over the FFT bands, while the compression is done by simply taking the logarithm of the smoothed power in each band. The overall shape of the power spectrum is further flattened by performing sliding window normalization of the smoothed log powers over the FFT bands. Next, the pitches of the current reference and degraded frame are computed using a stochastic subharmonic pitch algorithm described in chapter 4 of [21]. The ratio of the reference to degraded pitch ration is then used to determine a range of possible warping factors. If possible, this search range is extended by using the pitch ratios for the preceding and following frame pair.

The frequency alignment algorithm then iterates through the search range and warps the degraded power spectrum with the warping factor of the current iteration and processes the warped power spectrum using the preprocessing steps described above. The correlation of the processed reference and processed warped degraded spectrum is then computed for bins below 1500 Hz. After complete iteration through the search range, the "best" (i.e., that resulted in the highest correlation) warping factor is retrieved. The correlation of the processed reference and best warped degraded spectra is then compared against the correlation of the original processed reference and degraded spectra. The "best" warping factor is then kept if the correlation increases by a set threshold. If necessary, the warping factor is limited by a maximum relative change to the warping factor determined for the previous frame pair.

After the dewarping that may be necessary for aligning the frequency axis of reference and degraded signals, the frequency scale in Hz is warped toward the pitch scale in Bark, reflecting that at low frequencies, the human hearing system has a finer frequency resolution than at high frequencies. This is implemented by binning FFT bands and summing the corresponding powers of the FFT bands with a normalization of the summed parts. The warping function that maps the frequency scale in Hertz to the pitch scale in Bark approximates the values given in the literature [20]. The resulting reference and degraded signals are known as the pitch power densities $PPX(f)_n$ and $PPY(f)_n$ with f the frequency in Bark and the index n representing the frame index.

2.3 Computation of the Speech Active, Silent, and Super Silent Frames

POLQA operates on three classes of frames:

- Speech active frames where the frame level of the reference signal is above a level that is about 20 dB below the average;
- Silent frames where the frame level of the reference signal is below a level that is about 20 dB below the average; and
- Super silent frames where the frame level of the reference signal is below a level that is about 35 dB below the average level.

2.4 Calculation of the Frequency, Noise, and Reverb Indicators

The global impact of frequency response distortions, noise, and room reverberations is quantified separately. For the impact of overall global frequency response distortions, an indicator is calculated from the average spectra of reference and degraded signals. In order to make the estimate of the impact of frequency response distortions independent of additive noise, the average spectrum density of the degraded signal over the silent frames of the reference signal is subtracted from the pitch loudness density of the degraded signal. The resulting pitch loudness density of the degraded signal and the pitch loudness density of the reference signal are then averaged in each Bark band over all speech active frames for the reference and degraded signal. The difference in pitch loudness density between these two densities is then integrated over the pitch to derive the indicator for quantifying the impact of frequency response distortions (FREQ).

For the impact of additive noise, an indicator is calculated from the average spectrum of the degraded signal over the silent frames of the reference signal. The difference between the average pitch loudness density of the degraded signal over the silent frames and a zero reference pitch loudness density determines a noise loudness density

function that quantifies the impact of additive noise. This noise loudness density function is then integrated over the pitch to derive an average noise impact indicator (NOISE). This indicator is thus calculated from an ideal silence so that a transparent chain that is measured using a noisy reference signal will thus not provide the maximum MOS score in the final POLQA end-to-end speech quality measurement.

For the impact of room reverberations, the energy over time function (ETC) is calculated from the reference and degraded time series. The ETC represents the envelope of the impulse response *h(t)* of the system *H(f)*, which is defined as

$$Y_a(f) = H(f) \cdot X(f) \tag{1}$$

where $Y_a(f)$ is the spectrum of a level aligned representation of the degraded signal and $X(f)$ the spectrum of the reference signal. The level alignment is carried out to suppress global and local gain differences between the reference and degraded signal. The impulse response *h(t)* is calculated from *H(f)* using the inverse discrete Fourier transform. The ETC is calculated from the absolute values of *h(t)* through normalization and clipping. Based on the ETC up to three reflections are searched. In a first step the loudest reflection is calculated by simply determining the maximum value of the ETC curve after the direct sound. In the POLQA model, direct sound is defined as all sounds that arrive within 60 ms. Next a second loudest reflection is determined over the interval without the direct sound and without taking into account reflections that arrive within 100 ms from the loudest reflection. Then the third loudest reflection is determined over the interval without the direct sound and without taking into account reflections that arrive within 100 ms from the loudest and second loudest reflection. The energies and delays of the three loudest reflections are then combined into a single reverb indicator (REVERB).

2.5 Global and Local Scaling of the Reference Signal toward the Degraded Signal

The reference signal is now at the internal ideal level, i.e., about 73 dB SPL equivalent, while the degraded signal is represented at a level that coincides with the playback level. Before a comparison is made between the reference and degraded signal, the global level difference is compensated. Furthermore, small changes in local level are partially compensated to account for the fact that small enough level variations are not noticeable to subjects in a listening-only situation. The global level equalization is carried out on the basis of the average power of reference and degraded signal using the frequency components between 400 and 3500 Hz. The reference signal is globally scaled toward the degraded signal and the impact of the global playback level difference is thus maintained at this stage of processing. Similarly, for slowly varying gain distortions a local scaling is carried out for level changes up to about 3 dB using the full bandwidth of both the reference and degraded speech file.

2.6 Partial Compensation of the Original Pitch Power Density for Linear Frequency Response Distortions

In order to correctly model the impact of linear frequency response distortions, induced by filtering in the system under test, a partial compensation approach is used. To model the imperceptibility of moderate linear frequency response distortions in the subjective tests, the reference signal is partially filtered with the transfer characteristics of the system under test. This is carried out by calculating the average power spectrum of the original and degraded pitch power densities over all speech active frames. Per Bark bin, a partial compensation factor is calculated from the ratio of the degraded spectrum to the original spectrum.

2.7 Modeling of Masking Effects, Calculation of the Pitch Loudness Density Excitation

Masking is modeled by calculating a smeared representation of the pitch power densities. Both time- and frequency-domain smearing are taken into account (see Fig. 5). The time-frequency domain smearing uses the convolution approach as given in [22]. From this smeared representation, the representations of the reference and degraded pitch power density are re-calculated suppressing low amplitude time-frequency components, which are partially masked by neighboring loud components in the time-frequency plane. This suppression is implemented in two different manners, a subtraction of the smeared representation from the non-smeared representation and a division of the non-smeared representation by the smeared representation. The resulting, sharpened, representations of the pitch power density are then transformed to pitch loudness density representations using a modified version of Zwicker's power law [20]:

$$LX(f)_n = SL * \left(\frac{P_0(f)}{0.5}\right)^{0.22 * f_B * P_{fn}} * \left[\left(0.5 + 0.5\frac{PPX(f)_n}{P_0(f)}\right)^{0.22 * f_B * P_{fn}} - 1\right] \tag{2}$$

with *SL* the loudness scaling factor, $P_0(f)$ the absolute hearing threshold, f_B and P_{fn} a frequency- and level-dependent correction defined by:

$$\begin{array}{lll} f_B = -0.03 * f + 1.06 & for & f < 2.0 \quad Bark \\ f_B = 1.0 & for & 2.0 \le f \le 22 \quad Bark \\ f_B = -0.2 * (f - 22.0) + 1.0 & for & f > 22.0 \quad Bark \\ P_{fn} = (PPX(f)_n + 600)^{0.008} & & \end{array} \tag{3}$$

with *f* representing the frequency in Bark, $PPX(f)_n$ the pitch power density in frequency time cell *f, n.* The resulting two-dimensional arrays $LX(f)_n$ and $LY(f)_n$ are called pitch loudness densities.

2.8 Global Low Level Noise Suppression in Reference and Degraded Signals

Low levels of noise in the reference signal, which are not affected by the system under test (e.g., a transparent system) will be attributed to the system under test by subjects due to the absolute category rating test procedure. These low levels of noise thus have to be suppressed in

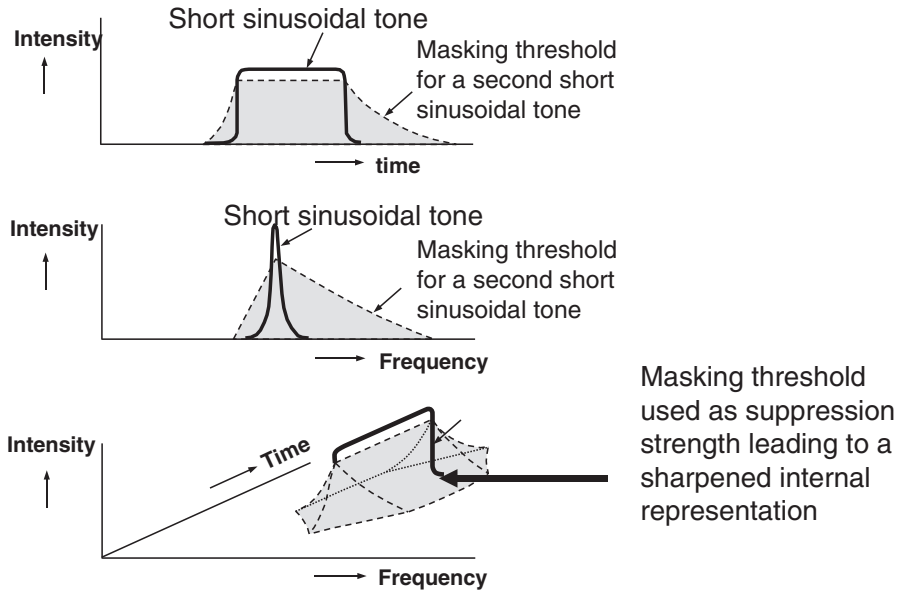

Fig. 5. Masking approach used in the POLQA perceptual model. The tent-like figure in the time-frequency plane is used to suppress the loudness of components of the signal.

the calculation of the internal representation of the reference signal. This "idealization process" is carried out by calculating the average steady state noise loudness density of the reference signal $LX(f)_n$ over the super silent frames as a function of pitch. This average noise loudness density is then partially subtracted from all pitch loudness density frames of the reference signal. The result is an idealized internal representation of the reference signal.

Steady state noise that is audible in the degraded signal has a lower impact than non-steady state noise. This holds for all levels of noise and the impact of this effect can be modelled by partially removing steady state noise from the degraded signal. This is carried out by calculating the average steady state noise loudness density of the degraded signal $LY(f)_n$ frames for which the corresponding frame of the reference signal is classified as super silent, as a function of pitch. This average noise loudness density is then partially subtracted from all pitch loudness density frames of the degraded signal. The partial compensation uses a different strategy for low and high levels of noise. For low levels of noise the compensation is only marginal while the suppression that is used becomes more aggressive for loud additive noise. The result is an internal representation of the degraded signal with an additive noise that is adapted to the subjective impact as observed in listening tests using an idealized, noise-free representation of the reference signal.

2.9 Local Scaling of the Distorted Pitch Loudness Density for Time-Varying Gain between Degraded and Reference Signal

Slow variations in gain are inaudible and small changes are already compensated for in the calculation of the reference signal representation (see Section 2.5). The remaining compensation, necessary before the correct internal representation can be calculated, is carried out in two steps: first the reference is compensated for signal levels where the degraded signal loudness is less than the reference signal loudness, and second the degraded is compensated for signal levels where the reference signal loudness is less than the degraded signal loudness. The first compensation scales the reference signal toward a lower level for parts of the signal where the degraded shows a severe loss of signal, such as in time clipping situations. The scaling is such that the remaining difference between reference and degraded represents the impact of time clips on the perceived local speech quality. Parts where the reference signal loudness is less than the degraded signal loudness are not compensated and thus additive noise and loud clicks are not compensated in this first step.

The second compensation scales the degraded signal toward a lower level for parts of the signal where the degraded signal shows clicks and for parts of the signal where there is noise in the silent intervals. The scaling is such that the remaining difference between reference and degraded represents the impact of clicks and slowly changing additive noise on the perceived local speech quality. While clicks are compensated in both the silent and speech active parts, the noise is compensated only in the silent parts.

2.10 Partial Compensation of the Original Pitch Loudness Density for Linear Frequency Response Distortions

Imperceptible linear frequency response distortions were already compensated by partially filtering the reference signal in the pitch power density domain. In order to further correct for the fact that linear distortions are less objectionable than non-linear distortions, the reference signal is now partially filtered in the pitch loudness domain. This is carried out by calculating the average loudness spectrum of the original and degraded pitch loudness densities over

all speech active frames. Per Bark bin, a partial compensation factor is calculated from the ratio of the degraded loudness spectrum to the original loudness spectrum. This partial compensation factor is used to filter the reference signal with a smoothed, lower amplitude, version of the frequency response of the system under test. After this filtering, the difference between the reference and degraded pitch loudness densities that result from linear frequency response distortions is diminished to a level that represents the impact of linear frequency response distortions on the perceived speech quality.

2.11 Final Scaling and Noise Suppression of the Pitch Loudness Densities

Up to this point, all calculations on the signals are carried out on the playback level as used in the subjective experiment. For low playback levels, this will result in a low difference between reference and degraded pitch loudness densities and in general in a far too optimistic estimation of the listening speech quality. In order to compensate for this effect, the degraded signal is now scaled toward a "virtual" fixed internal level. After this scaling, the reference signal is scaled toward the degraded signal level and both the reference and degraded signal are now ready for a final noise suppression operation. This noise suppression takes care of the last parts of the steady-state noise levels in the loudness domain that still have a too big impact on the speech quality calculation. The resulting signals are now in the perceptually relevant internal representation domain and from the ideal pitch-loudness-time $LX_{ideal}(f)_n$ and degraded pitch-loudness-time $LY_{deg}(f)_n$ functions the so-called disturbance densities can be calculated. Four different variants of the ideal and degraded pitch-loudness-time functions are calculated, two variants focused on the disturbances for normal and big distortions, and two focused on the added disturbances for normal and big distortions.

2.12 Calculation of the Final Disturbance Densities

Two different flavors of the disturbance densities are calculated. The first one, the "normal" disturbance density, is derived from the difference between the ideal pitch-loudness-time $LX_{ideal}(f)_n$ and degraded pitch-loudness-time $LY_{deg}(f)_n$ functions. The second one is derived from the ideal pitch-loudness-time and the degraded pitch-loudness-time function using versions that are optimized with regard to introduced degradations and is called added disturbance. In this added disturbance calculation, signal parts where the degraded power density is larger than the reference power density are weighted with a factor dependent on the power ratio in each pitch-time cell, the asymmetry factor.

In order to be able to deal with a large range of distortions, two different versions of the processing are carried out, one focused on small to medium distortions and one focused on medium to big distortions. The switching between the two is carried out on the basis of a first estimation from the disturbance focused on small to medium levels of distortions. This processing approach leads to the necessity of calculating four different ideal pitch-loudness-time functions and four different degraded pitch-loudness-time functions in order to be able to calculate a single disturbance and a single added disturbance function (see Fig. 3) that are then compensated for a number of different types of severe amounts of specific distortions.

Severe deviations from the optimal listening level are quantified by an indicator directly derived from the active speech level of the degraded signal. This global indicator (LEVEL) is also used in the calculation of the MOS-LQO.

Severe distortions introduced by frame repeats are quantified by an 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.

Severe deviations from the "ideal" timbre of the degraded signal are quantified by an indicator derived from the difference in loudness between an upper frequency band and a lower frequency band. A timbre indicator is calculated from the difference in loudness in the Bark bands between 2 and 12 Bark in the low frequency part and 7–17 Bark in the upper range (i.e., using a 5 Bark overlap) of the degraded signal that "punishes" any severe imbalances irrespective of the fact that this could be the result of an incorrect voice timbre of the reference speech file. Compensations are carried out for each frame individually as well as on a global level. This compensation also has an impact when measuring the quality of devices that are transparent. When reference signals are used that show a significant deviation from the "ideal" timbre, the system under test will be judged as non-transparent even if the system does not introduce any degradation into the reference signal.

The impact of severe frequency domain peaks in the disturbance loudness function is quantified in the FLATNESS indicator, which is also used in the calculation of the MOS-LQO.

Severe noise level variations that focus the attention of subjects toward the noise are quantified by a noise contrast indicator derived from the degraded signal frames for which the corresponding reference signal frames are silent.

Severe jumps in the alignment are detected in the alignment and the impact is quantified by a compensation factor.

Finally, the disturbance and added disturbance densities are clipped to a maximum level and the variance of the disturbance and the jumps in the loudness are used to compensate for specific time structures of the disturbances.

2.13 Aggregation of the Disturbance over Pitch, Spurts, and Time, Mapping to the Intermediate MOS Score

The final disturbance $D(f)_n$ and added disturbance $DA(f)_n$ densities are integrated per frame over the pitch axis resulting in two different disturbances per frame, one derived from the disturbance and one derived from the added

disturbance, using an L_1 integration (see Fig. 4):

$$D_n = \sum_{f=1,..Number\ of\ Barkbands} |D(f)_n|\ W_f$$

$$DA_n = \sum_{f=1,..Number\ of\ Barkbands} |DA(f)_n|\ W_f \quad (4)$$

with W_f a series of constants proportional to the width of the Bark bins.

Next these two disturbances per frame are averaged over a concatenation of six consecutive speech frames, defined as a speech spurt, with an L_4 and an L_1 weighting for the disturbance and for the added disturbance, respectively.

$$DS_n = \sqrt[4]{\frac{1}{6}\sum_{m=n,..n+6} D_m^4}$$

$$DAS_n = \frac{1}{6}\sum_{m=n,..n+6} D_m \quad (5)$$

Finally a disturbance and an added disturbance are calculated per file from an L_2 averaging over time:

$$D = \sqrt[2]{\frac{1}{N}\sum_{n=1,..number\,Of\,Frames} DS_n^2}$$

$$DA = \sqrt[2]{\frac{1}{N}\sum_{n=1,..number\,Of\,Frames} DAS_n^2}$$

$$N = Number\ of\ frames \quad (6)$$

The added disturbance is compensated for loud reverberations and loud additive noise using the REVERB and NOISE indicators. The two disturbances are then combined with the frequency indicator (FREQ) to derive an internal indicator that is linearized with a third order regression polynomial to get a MOS like intermediate indicator.

2.14 Computation of the Final POLQA MOS-LQO

The raw POLQA score is derived from the MOS like intermediate indicator using four different compensations:

- Two compensations for specific time-frequency characteristics of the disturbance, one calculated with an L_{511} aggregation over frequency, spurts and time, and one calculated with an L_{313} aggregation over frequency, spurts and time;
- One compensation for very low presentation levels using the LEVEL indicator;
- One compensation for big timbre distortions using the FLATNESS indicator in the frequency domain.

The training of this mapping is carried out on a large set of degradations, including degradations that were not part of the POLQA benchmark. These raw MOS scores are for the major part already linearized by the third order polynomial mapping used in the calculation of the MOS like intermediate indicator (Section 2.13).

Finally the raw POLQA MOS scores are mapped toward the MOS-LQO scores using a third order polynomial that is optimized for the 62 databases that were available in the final stage of the POLQA standardization. In narrowband mode the maximum POLQA MOS-LQO score is 4.5 while in super-wideband mode this point lies at 4.75. An important consequence of the idealization process is that under some circumstances, when the reference signal contains noise or when the voice timbre is severely distorted, a transparent chain will not provide the maximum MOS score of 4.5 in narrowband mode or 4.75 in super-wideband mode.

3 POLQA PERFORMANCE RESULTS

As stated before, the ITU-T Study Group 12 POLQA benchmark consisted of three phases. In the first phase 6 models were trained on 16 wideband/super-wideband databases that used diffuse field equalized headphone presentation and 29 narrowband databases that used standard telephone handset playback. In the validation phase all models had to predict the subjective scores of 8 wideband/super-wideband databases and 9 narrowband databases that were unknown to all candidates. The focus of POLQA was on the wideband/super-wideband databases allowing assessing all possible future high quality voice services. The narrowband training and validation was carried out in order to be able to have direct comparisons of the performance of POLQA P.863 with PESQ P.862 in the classical narrowband (300-3400Hz) telephone situation. A description of the subjective experiments is provided is Section 1 of this paper.

The results of both the training and validation were used to compare the performance of the six submitted models. The performance was measured in terms of the Root Mean Squared Error (RMSE) compensated for the 95% confidence interval of the MOS score obtained in the subjective test, i.e., the distance between the predicted MOS score and the closest point of the 95% confidence interval was used as the model prediction error. This method for calculating the RMSE, coined RMSE*, has the advantage that it takes into account the reliability of each MOS score, as opposed to a straight-forward correlation calculation for which this is not possible. In order to take into account the context of a subjective test, which may result in different MOS scores for the same degradation, a third order regression is applied for each speech quality database separately. An overview of the procedure for calculating the RMSE* is given in Appendix I.4 of ITU-T Recommendation P.863 [11].

With this RMSE* calculation the three best models showed a significantly better performance when compared to the remaining three models. These models from OPTICOM, SwissQual, and TNO were integrated into a single POLQA model that outperformed each of the underlying proposals in terms of RMSE*.

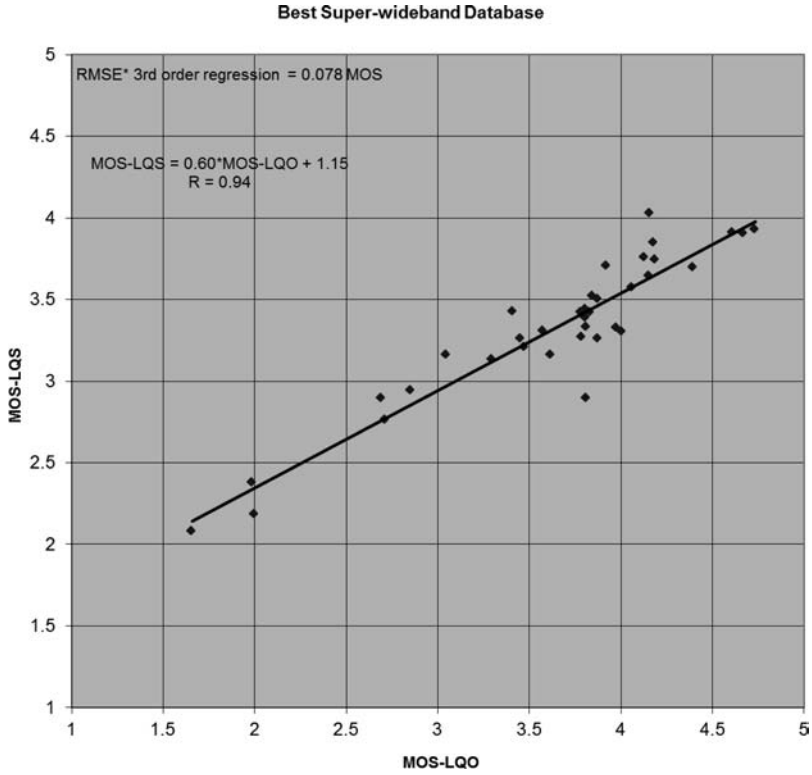

Fig. 6. POLQA result for the best case narrowband experiment using all phase 1 and phase 2 subjective tests. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results, is 0.06 MOS. The linear correlation is excellent (0.97) and close to the ideal regression Y = X.

The new POLQA model provides a significant improvement over PESQ in both the narrowband and (super)wide band mode. The average RMSE*, over all available 38 narrowband databases is 0.19 MOS for PESQ and 0.14 MOS for POLQA. The average corrected root mean square error over all available 7 wideband databases is 0.34 MOS for PESQ and 0.15 MOS for POLQA, showing the huge improvement of POLQA over PESQ. For all available 17 super-wideband data the corrected root mean square error for POLQA is 0.21 MOS, much better than the performance of PESQ with the wideband data (0.34 MOS). Of course a direct PESQ-POLQA comparison is not possible in super-wideband mode because the maximum sample rate for PESQ is 16 kHz. Six examples of the performance of POLQA are given in Figs. 6 through 11. They show the best and worst case performance in terms of RMSE*, using narrowband, wideband, and super-wideband databases as available from all data used in the benchmark. Besides the 3rd order RMSE* the more traditional linear correlation results are also provided. The overall performance results of POLQA in terms of linear correlations are excellent, in the narrowband mode the average linear correlation is 0.94 with a worst case performance of 0.85, while in super-wideband mode these numbers are only slightly lower, 0.92 and 0.83 respectively, taking into account both wideband and super-wideband data. It should be noted that for the super-wideband mode the worst case database for correlation is not the same database as for RMSE*. A complete overview of the results on all 62 databases, including the comparison between POLQA and PESQ, is given in Appendix I.3 of ITU-T Recommendation P.863 [11].

In narrowband mode POLQA is technically backwards compatible with PESQ, they both model telephone narrowband listening with a maximum MOS value of 4.5. The wideband mode of PESQ has a maximum MOS of 4.5 while the maximum with POLQA in wideband/super-wideband mode is 4.75, making it difficult to compare the scores. Also the performance of PESQ is too low to allow for direct comparisons between PESQ and POLQA in wideband/super-wideband mode.

Another important difference between PESQ and POLQA is the fact that the maximum MOS value with POLQA is not automatically achieved for a transparent speech link while for PESQ such a link will always provide the maximum MOS. This is due to the idealization process as used in the POLQA perceptual model that will compensate for non-optimal voice timbres and possible residual noise in the reference recording. If a non-optimal reference recording is provided to POLQA as both the reference and the degraded, the idealization will lead to a difference in the internal representation.

4 FINAL BENCHMARKING AND COMPARISON IN PERFORMANCE BETWEEN PESQ AND POLQA

The final benchmarking of the POLQA algorithm will be carried out in the field, but it should be noted that

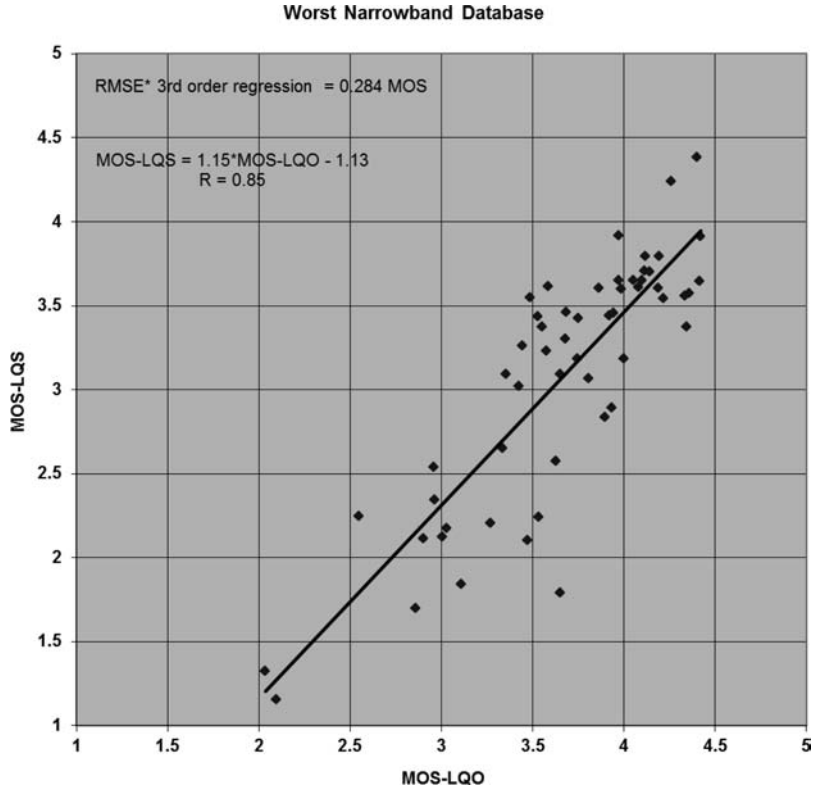

Fig. 7. POLQA result for the worst case narrowband experiment using all phase 1 and phase 2 subjective tests. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results, is 0.28 MOS. The linear correlation is very good (0.85) and close to the ideal regression Y = X.

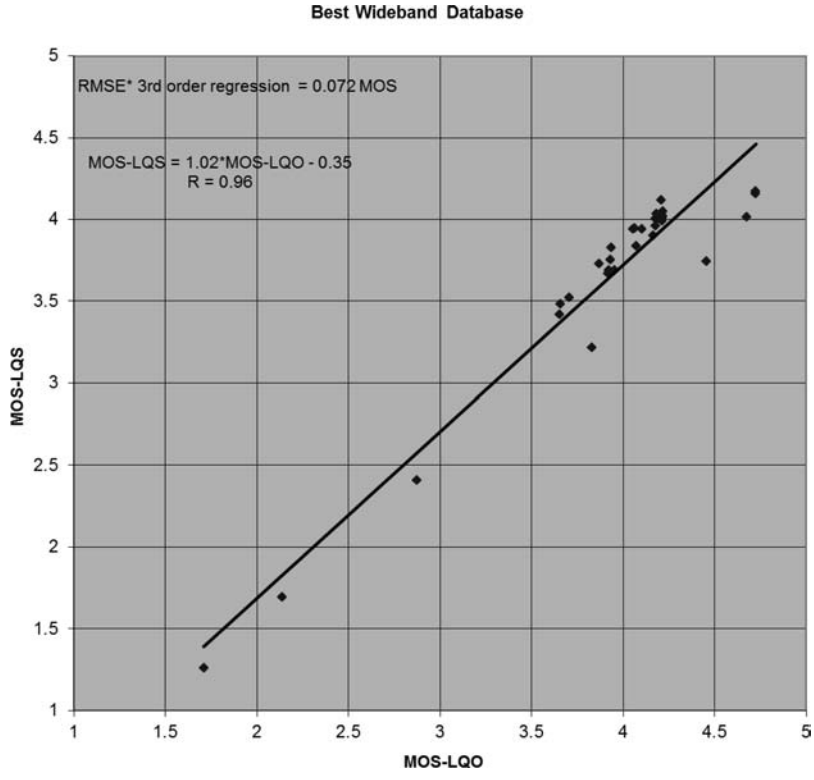

Fig. 8. POLQA result for the best case wideband experiment using all phase 1 and phase 2 subjective tests. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results, is 0.07 MOS. The linear correlation is excellent (0.96) and very close to the ideal regression Y = X.

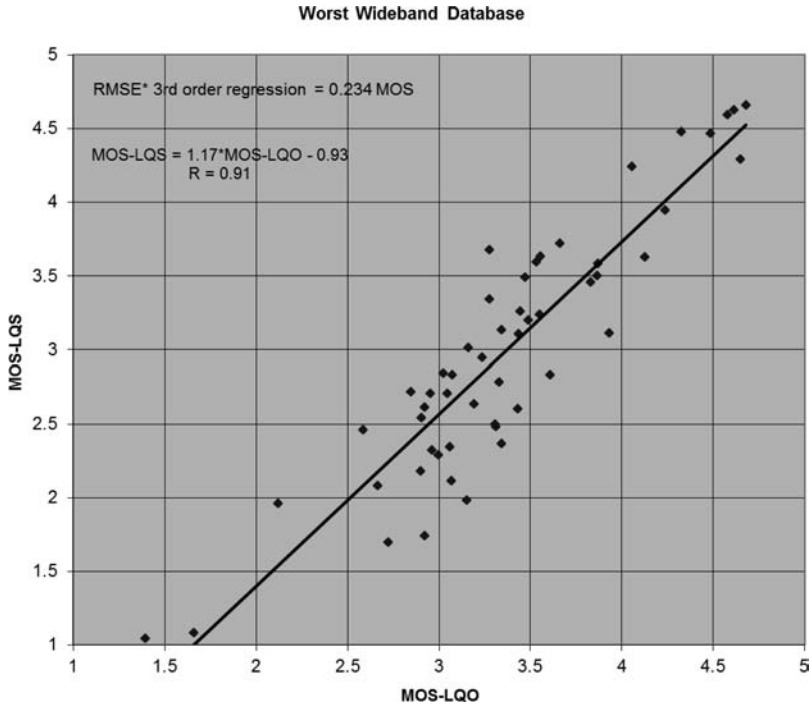

Fig. 9. POLQA result for the worst case wideband experiment using all phase 1 and phase 2 subjective tests. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results, is 0.23 MOS. The linear correlation is excellent (0.91) and close to the ideal regression Y = X.

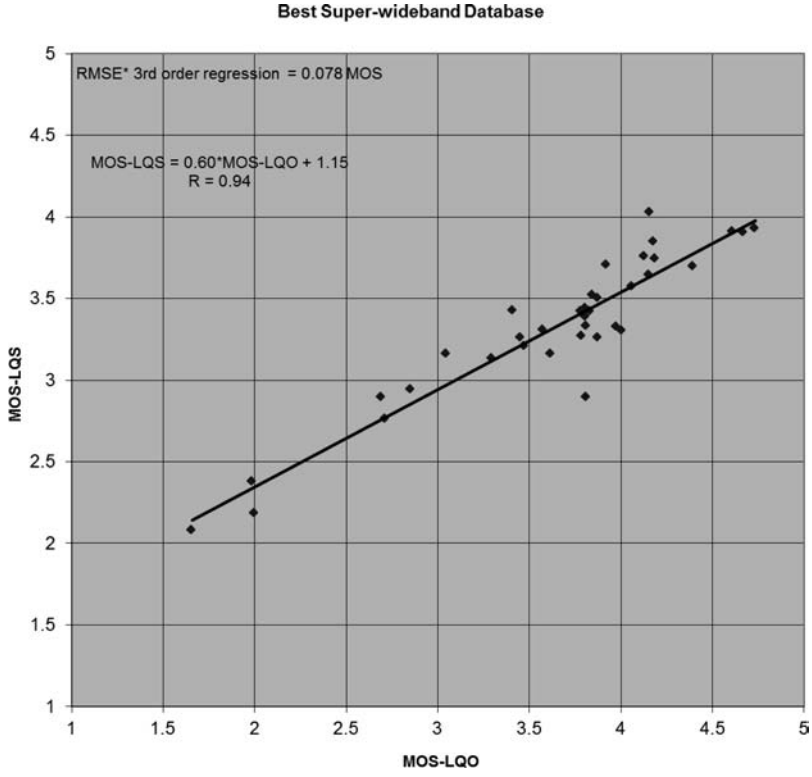

Fig. 10. POLQA result for the best case super-wideband experiment using all phase 1 and phase 2 subjective tests. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results, is 0.08 MOS. The linear correlation is excellent (0.94), the significant deviation from ideal regression Y = X is caused by an imbalance in the subjective test that did not contain enough sever degradations with low subjective MOS scores.

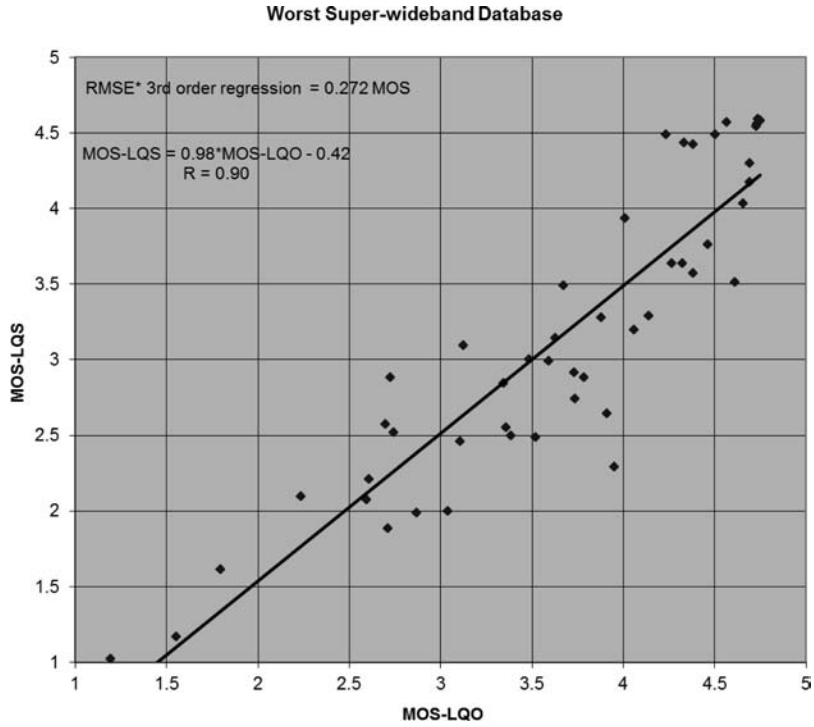

Fig. 11. POLQA result for the worst case super-wideband experiment using all phase 1 and phase 2 subjective tests. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results, is 0.27 MOS. The linear correlation is excellent (0.91) and very close to the ideal regression Y = X.

the underlying algorithms from OPTICOM, SwissQual, and TNO were already validated on unknown databases during the second phase of the ITU benchmarking where all models were submitted to the ITU and new databases were created to validate model performance on unknown databases. The best underlying models used in the final POLQA standard only showed a marginal increase in the RMSE* from less than 0.20 MOS on the training set in phase 1 (45 databases more then 28,000 speech files) to less than 0.23 MOS on the unknown validation set in phase 2 (17 databases, more than 17,000 speech files). The final integrated model as developed in phase 3 was trained on all POLQA benchmark databases, i.e., phase 1 training and phase 2 validation data, using the same general model stability approach towards behavior on unknown data as used in the best underlying models. This final model outperformed each of the three underlying models from which it was built using all the 62 databases of phase 1 and 2 in the comparison. The total amount of speech files, more than 45,000 were used, guarantees that the model is not over trained.

In order to further check the model for overtraining, and to further compare the results of POLQA SWB with PESQ WB, three databases that were not used in the final POLQA model training were used in a final validation of the phase 3 POLQA model. These databases used 16 kHz sampled wideband recordings of the degraded speech, allowing to run PESQ WB on them. Running POLQA in SWB mode then allows a direct comparison of PESQ and POLQA because possible bias effects resulting from a difference in MOS scale training is compensated by using a third order regression comparison. Due to the fact that from the 24 wideband/super-wideband databases used in the phase 3 model training only seven databases were wideband, it is expected that this is the weakest point of the POLQA model. In the final model validation and model comparison the following three databases were used.

- A wideband database with narrowband speech and wideband background noise resulting from a mobile handset evaluation using electric coupling into the handset and acoustic recording of the loudspeaker output in the presence of background noise with a Head and Torso Simulator (Fig. 12).
- A wideband database with narrowband speech and wideband background noise resulting from a mobile handset evaluation using acoustic coupling into one handset and acoustic recording at a second (Fig. 13).
- A database with artificially generated time clipping, frequency response distortion and background noise using extreme degradations (Fig. 14).

The processing results of these three databases show that the final model generalizes well toward distortions not used in the training and also shows that POLQA is significantly better than PESQ, although the improvement is less than found in the databases of phase 1 and 2, where the average RMSE* dropped from 0.34 to 0.15, while in this validation the average RMSE* dropped from 0.28 to 0.18. In terms of

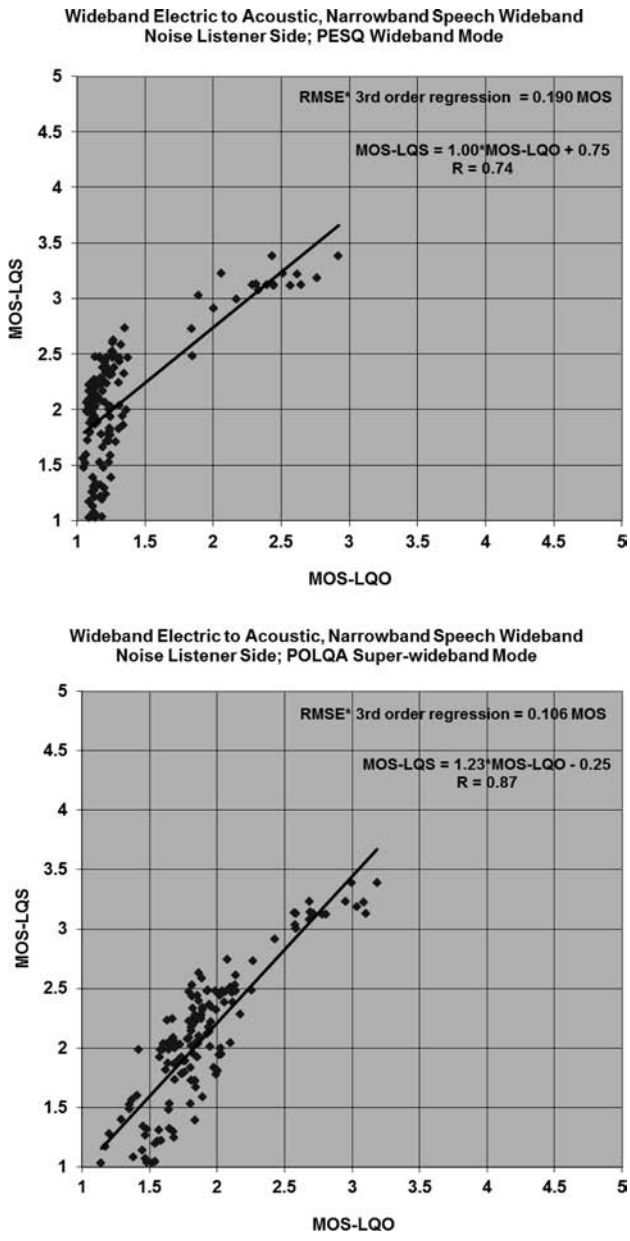

Fig. 12. Comparison of PESQ and POLQA results for the wideband electric to acoustic database with narrowband speech and wideband background noise. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results, drops from 0.19 down to 0.11 MOS, while the linear correlation increases from 0.74 to 0.87.

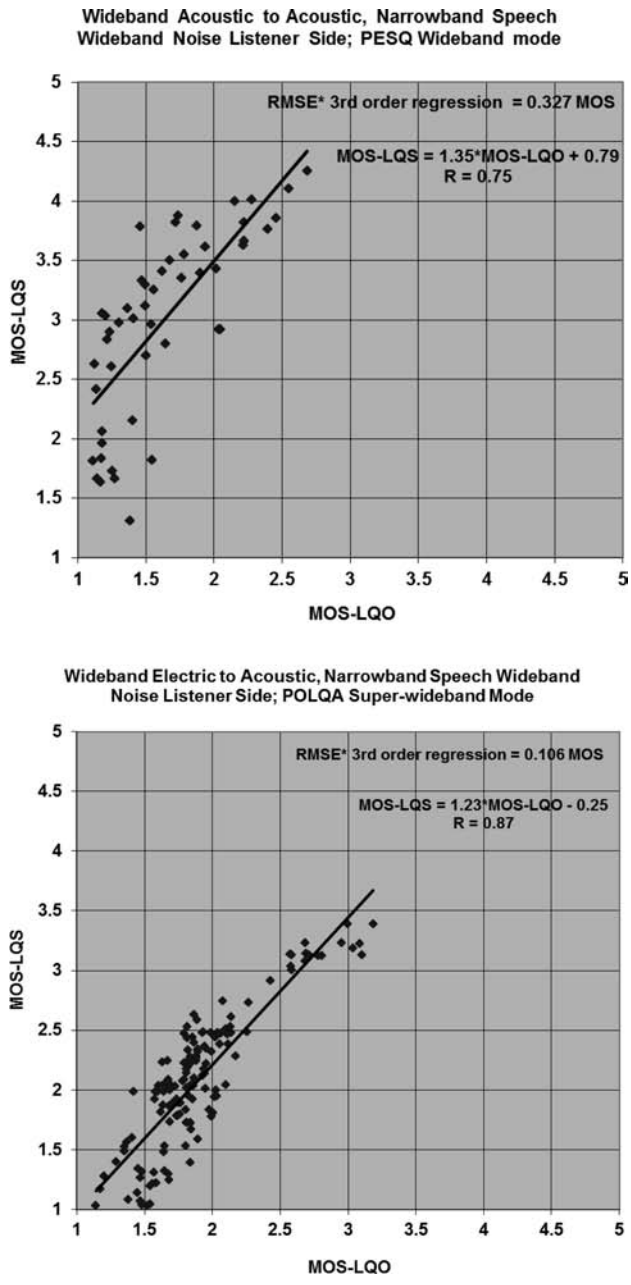

Fig. 13. Comparison of PESQ and POLQA results for the wideband acoustic to acoustic database with narrowband speech and wideband background noise. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results, drops from 0.33 down to 0.19 MOS, while the linear correlation increases from 0.75 to 0.87.

linear correlation the average performance increases from 0.76 for PESQ to 0.87 for POLQA.

5 CONCLUSIONS

This paper deals with the perceptual model of the new objective speech quality assessment method that resulted from a benchmark carried out by the ITU-T (International Telecommunication Union, Telecom sector) in order to define a technology update of PESQ (ITU-T Recommendation P.862), the established worldwide industry standard for the objective assessment of speech quality. Together with the temporal alignment provided in Part I, it gives a full description of the new ITU-T Recommendation P.863 POLQA. The benchmark carried out within the ITU-T showed that from six candidate algorithms that were originally submitted, three fulfilled the ITU-T requirements and were thus selected for the final standardization. These models showed excellent overall performance and excellent stability in predicting subjective scores for distortions not used in the model training. In order to derive a unique

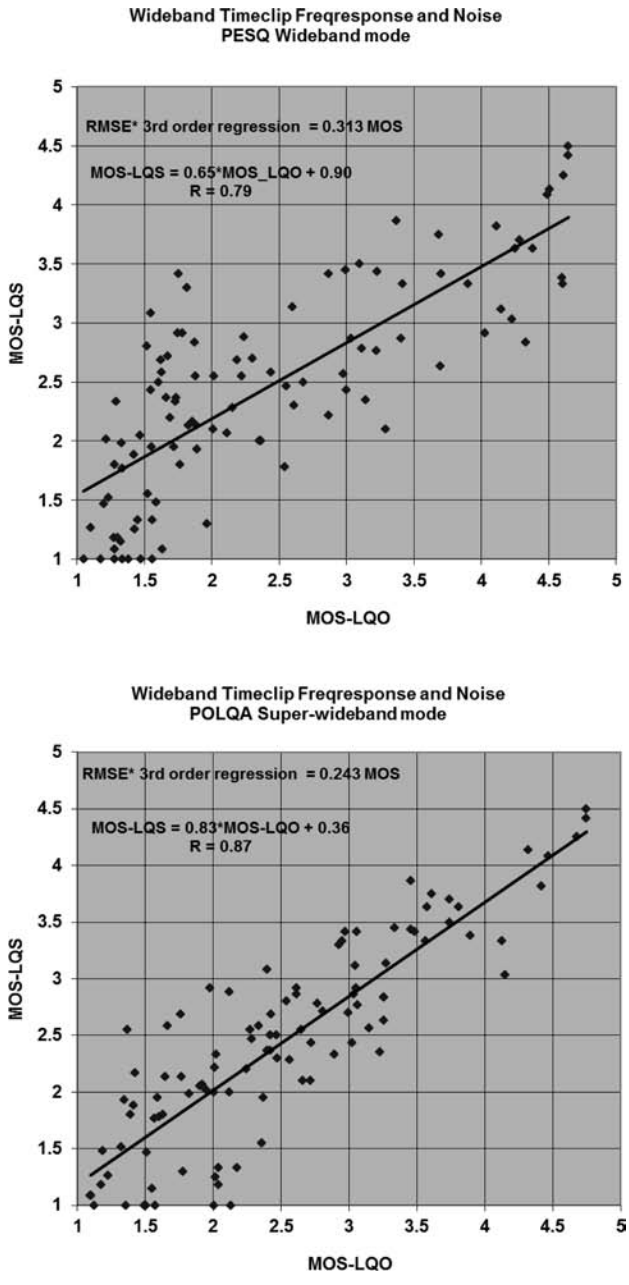

Fig. 14. Comparison of PESQ and POLQA results for the artificially generated time clipping, frequency response distortion, and background noise. The average root mean square error, corrected for the 95% confidence interval (RMSE*), using a 3rd order regression between subjective and objective results drops from 0.31 down to 0.24 MOS for PESQ and POLQA respectively, while the linear correlation increases from 0.79 to 0.87.

measurement algorithm these models from OPTICOM, SwissQual, and TNO were integrated into a single joint model under the name POLQA (Perceptual Objective Listening Quality Assessment). POLQA provides a significant improvement over PESQ for narrowband (300–3400 Hz) as well as for wideband (50–7000 Hz) speech quality measurement. Also, POLQA allows quality assessment using super-wideband (20–14000 Hz) speech signals. It was accepted as new ITU-T Recommendation P.863 in January 2011.

POLQA outperforms PESQ in assessing any kind of degradation, making it an ideal tool for all speech quality measurements from low end to HD voice communication in today's and future Voice-over-IP based and mobile networks.

POLQA is also suited for distortions that are outside the scope of PESQ, such as linear frequency response distortions, time stretching/compression as found in Voice-over-IP, certain types of codec distortions, and reverberations. Furthermore POLQA is able to assess the impact of playback volume in the range between 53 and 78 dB(A), thus allowing system designers to assess the impact of the fact that users adapt their playback volume toward the background noise level.

Finally, users of POLQA should be aware of the fact that a transparent chain for which the quality is measured with poorly recorded reference signals, containing too much noise and/or an incorrect voice timbre, will not provide the maximum MOS score as one would expect from a transparent voice link. POLQA thus only provides reliable speech quality predictions when high quality reference voice recordings are used in the objective assessment. A test to see whether a voice recording is close enough to the ideal can be carried out by simply providing the voice recording as degraded and reference signal to POLQA. The internal idealization process carried out in the reference chain of POLQA should lead to an internal representation that is close to the internal representation of the reference as processed in the degraded chain of the POLQA algorithm, leading to the maximum MOS score of 4.75 for a transparent voice link.

6 REFERENCES

[1] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," International Telecommunication Union, Helsinki (1993); revised Geneva, Switzerland (1996).

[2] ITU-T Rec. P.830, "Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs," International Telecommunication Union, Geneva, Switzerland (1996 Feb.).

[3] ITU-T Rec. P.861, "Objective Quality Measurement of Telephone Band (300–3400 Hz) Speech Codecs," International Telecommunication Union, Geneva, Switzerland (1996 Aug.).

[4] J. G. Beerends and J. A. Stemerdink, "A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 42, pp. 115–123 (1994 Mar.).

[5] ITU-T Rec.P.862, "Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," International Telecommunication Union, Geneva, Switzerland (2001 Feb.).

[6] ITU-T Rec. P.862.1, "Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO," International Telecommunication Union, Geneva, Switzerland (2003 Nov.).

[7] A. W. Rix, M. P. Hollier, A. P. Hekstra and J.G. Beerends, "PESQ, the New ITU Standard for Objective Measurement of Perceived Speech Quality, Part I – Time Alignment," *J. Audio Eng. Soc.*, vol. 50, pp. 755–764 (2002 Oct.).

[8] J .G. Beerends, A. P. Hekstra, A. W. Rix, and M.P. Hollier, "PESQ, the New ITU Standard for Objective Measurement of Perceived Speech Quality, Part II – Perceptual Model," *J. Audio Eng. Soc.*, vol. 50, pp. 765–778 (2002 Oct.).

[9] ITU-T Rec. P.862.2, "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," International Telecommunication Union, Geneva, Switzerland (2005 Nov.).

[10] B. C. Bispo, P. A. A. Esquef, L. W. P. Biscainho et. al., "EW-PESQ: A Quality Assessment Method for Speech Signals Sampled at 48 kHz," *J. Audio Eng. Soc.* vol. 58, pp. 251–268 (2010 Apr.).

[11] ITU-T Rec. P.863, "Perceptual Objective Listening Quality Assessment," Geneva, Switzerland (2011 Jan.).

[12] ITU-T Rec. P.56, "Objective Measurement of Active Speech Level," Geneva (1993 Mar.).

[13] ITU-T Rec. P.810, "Modulated Noise Reference Unit (MNRU)" (1996 Feb.).

[14] ITU-T Rec. P.50, "Artificial Voices," Geneva (1999 Mar.).

[15] D. F. Hoth, "Room Noise Spectra at Subscribers' Telephone Locations," *J. Acoust. Soc. Am.*, vol. 12, pp. 499–504 (1941 Apr.).

[16] L. Malfait and J. Berger, "Analysis of the POLQA Full-Scale Super-Wideband Experiments," ITU-T Contribution COM 12-20, Geneva(2009 Oct.).

[17] J. G. Beerends and J. A. Stemerdink, "Modelling a Cognitive Aspect in the Measurement of the Quality of Music Codecs," presented at the 96th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 42, p. 394 (1994 May), convention paper 3800.

[18] J. G. Beerends, B. P. Busz, P. Oudshoorn, J. M. Van Vugt, O. K. Ahmed, and O. A. Niamut, "Degradation Decomposition of the Perceived Quality of Speech Signals on the Basis of a Perceptual Modeling Approach," *J. Audio Eng. Soc.*, vol. 55, pp. 1059–1076 (2007 Dec.).

[19] B. C. J. Moore, B. R. Glasberg, C. J. Plack, and A. K. Biswas, "The Shape of the Ear's Temporal Window," *J. Acoust. Soc. Am.*, vol. 83, pp. 1102–1116 (1988 Mar.).

[20] E. Zwicker, R. Feldtkeller, *Das Ohr als Nachrichtenempfänger* (S. Hirzel Verlag, Stuttgart, 1967).

[21] J. G. Beerends, "Pitches of Simultaneous Complex Tones, Chapter 5: A Stochastic Subharmonic Pitch Model," Ph.D. dissertation, Technical University of Eindhoven (April 1989).

[22] J. G. Beerends and J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 40, pp. 963–978 (1992 Dec.).

See Part I of this paper for the bios of the authors.