Subjective and Objective Assessment of Full Bandwidth Speech Quality

John G. Beerends, Niels M. P. Neumann, Egon L. van den Broek, Anna Llagostera Casanovas, Jovana Torres Menendez, Christian Schmidmer, and Jens Berger

***Abstract*—With the introduction of fullband speech coding the question arises what role frequency components above 14 kHz play in speech quality assessment. On the one hand, our results show that bandwidth limitation from 24 kHz down to 14 kHz is not audible to even the most critical subject. On the other hand, 14–24 kHz band limited, audible levels of noise clearly decrease the perceived quality, especially for young subjects with healthy ears. Furthermore, modern high-quality voice links, using the latest speech codecs, often apply advanced buffering schemes that introduce a new type of audible degradation: micropauses. We investigated the impact of i) bandwidth limitation, ii) coding schemes, iii) micropause, and iv) noise on the perceived quality. Subjective results and objective predictions based on ITU-T recommendation P.863 POLQA are compared. For accurate prediction of the impact of micropauses and noise degradations small model adaptations are suggested. In contrast codec degradations and bandwidth limitation are already predicted with very high accuracy by POLQA: $r = 0.98$, RMSE* $= 0.05$ Mean Opinion Score (MOS).**

***Index Terms*—Full bandwidth speech quality, POLQA.**

I. Introduction

In THE last decade there has been a trend to extend the bandwidth used in speech coding standards from the classical 3.5 kHz narrowband (NB) via 7 kHz wideband (WB) to 14 kHz super wideband (SWB) and >20 kHz fullband (FB) [1]–[5]. Recent studies show the relevance of this trend [6]. In general, frequencies above 10 kHz can be important in speech perception [7], [8], where speech quality is degraded when the upper cutoff frequency is decreased below 16 kHz [9].

The fullband speech codecs [1] strive for excellent speech quality in telephony applications, but it is unclear to what extent frequency components above 14 kHz play a role in speech quality assessment. This is relevant for both codec developers and developers of objective speech quality assessment methods, who use perceptual models to mimic human perception. We know that speech contains little energy above 14 kHz [10]. However, a bandwidth extension from 14 to above 20 kHz may lead to an audible improvement in case of subtle effects in sharp onsets/offsets (i.e., transients). Furthermore, perceived speech quality declines when loud distortions or noise components are introduced above 14 kHz. Therefore, to correctly assess distortions above 14 kHz, speech quality assessment methods should include models for the behavior of the auditory system above 14 kHz.

Apart from bandwidth extension, distortions can also be introduced by coding schemes. Recently more cognitive related coding schemes have been developed [11]. These introduce new types of degradations. These new coding trends are reflected in the updating of the ITU-T recommendations for the objective assessment of speech quality, P.862 PESQ (Perceptual Evaluation of Speech Quality) [12]–[14] and P.863 POLQA (Perceptual Objective Listening Quality Prediction) [15]–[17].

Assessing the impact of bandwidth extensions have been considered before. In 2011 Ekman *et al.* showed that the impact of degradations in super wideband speech signals cannot be accurately predicted by the extension of PESQ towards wideband [18]. They proposed a new approach that allowed for an accurate prediction. In the same year Nunes *et al.* proposed to use a degradation classifier in the assessment of fullband speech quality [19]. Later in 2017, Abel *et al.* proposed a method to assess the speech quality of artificially bandwidth extended speech signals [20]. Other methods for speech and audio quality assessment are continuously being introduced and evaluated [21]–[24]. These methods are evaluated on ever more advanced speech processing techniques, however, sometimes with conflicting results.

The follow up recommendation of PESQ is ITU-T P.863 POLQA [16], [17] and has recently been updated [15]. This update allows to accurately predict the impact of degradations of advanced signal processing techniques, using super wideband speech signals. For fullband speech signals however, no validation of the correct assessment of degradations has been provided yet.

With the bandwidth extensions, the impact of subject's hearing capabilities (e.g., a decline due to aging) on perceived speech quality becomes progressively more important. For example, only young subjects with healthy ears can be used in the assessment of the impact of degradations above 10 kHz. On the other hand, the effect of age or (weak) hearing impairment also has

Manuscript received June 13, 2019; revised October 2, 2019 and November 20, 2019; accepted November 22, 2019. Date of publication December 6, 2019; date of current version December 24, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tan Lee. *(Corresponding author: John G. Beerends.)*

J. G. Beerends and N. M. P. Neumann are with the The Netherlands Organization for Applied Scientific Research, 2509 The Hague, The Netherlands (e-mail: john.beerends@tno.nl; niels.neumann@tno.nl).

E. L. van den Broek is with the Utrecht University, 3512 Utrecht, The Netherlands (e-mail: vandenbroek@acm.org).

A. Llagostera Casanovas and J. Berger are with the Rohde & Schwarz SwissQual AG, 4528 Zuchwil, Switzerland (e-mail: anna.llagostera@rohde-schwarz.com; jens.berger@rohde-schwarz.com).

J. T. Menendez and C. Schmidmer are with the OPTICOM GmbH, 91052 Erlangen, Germany (e-mail: jovanav@opticom.de; cs@opticom.de).

Digital Object Identifier 10.1109/TASLP.2019.2957871

an impact on the perceptual modelling as used in POLQA and other objective perceptual measurement methods.

Another trend in modern, high quality, speech coding are advanced buffering schemes. Such schemes deal with varying delays as found in voice over IP links. These advanced buffering schemes introduce micropauses in the speech signal which may lead to audible degradations.

The main goal of this paper is two-fold:

1) Firstly, a subjective test is set up to investigate the impact of:
   a) the quality of hearing of the subjects,
   b) audio bandwidth in relation to the quality of hearing, and,
   c) coding techniques.
2) Secondly, this paper tests the performance of POLQA for fullband speech signals degraded by modern voice link degradations [1]–[5], including micropauses [25]–[27].

The remaining paper is setup as follows: In Section II a description of the speech signals and the degradations used in the subjective test are provided. Section III presents the subjective test, i.e., the experimental setup, subject selection/training and the results of the subjective experiment. Section IV presents the basics of the objective measurement approach and the objective measurement results. A discussion of both the subjective and objective results is given in Section V and Section VI provides the final conclusions.

II. Source Material

The dataset consists of both reference and degraded speech files, with a subjective quality score for each degraded speech file. We used this dataset for both subjective and objective evaluation. To assess high quality coding systems, we used reference signals, recorded in a large anechoic room (6 × 6 × 6 m). A high-quality studio microphone (16 mm B&K 4003, 10 Hz – 20 kHz, equivalent noise level 17 dB (A)) was used in combination with a 16-bit, 48 kHz sampling standard AD converter (Creative Labs SB0490). Each reference speech file consists of two Dutch spoken sentences separated by a silent period of about 1.5 seconds. The minimum amount of active speech is about 3 seconds.

The digital Active Speech Level (ASL according to ITU-T P.56 [28]) of the signals is equalized to -26 dBov (dB overload) for presentation at the nominal level. The corresponding nominal Sound Pressure Level (SPL) in the acoustical domain is 73 dB(A) SPL at the ear reference point. A high-quality diffuse field equalized headphone is used (HEAD acoustics HPS IV amplifier/equalizer in combination with a Sennheiser HD600).

Minor residues of background noise are suppressed to optimize the quality of these clean reference recordings. Background noise might originate from breathing noise of the talker and/or from system noise from the microphone and AD converter. Noise suppression was performed following [29], where manual noise suppression in silent intervals slightly improved quality scores.

Although probably inaudible, inspection of the recordings showed relevant high frequency components up to 24 kHz. These clean fullband speech signals are defined as the unprocessed reference conditions. Eight different speech samples from four different talkers, two males and two females, are used in the experiment. In the experiment, we degraded these eight speech samples with twenty-four different degradations. Each degradation can be found in modern high-quality voice links:

- Bandwidth limitation to 14 kHz (SWB), 7 kHz (WB) or 3.5 kHz (NB). FFT filters are used with slopes of 60, 100 and 200 dB/kHz respectively
- Low and medium levels of pink noise (SNR = 40 dB and 30 dB respectively)
- Micropauses introduced in low level speech intervals in both sentences. These represent degradations introduced by packet switched networks (length micropause: 30 ms, 50 ms, 100 ms; single occurrence in each sentence)
- Micropauses introduced in medium level speech intervals in both sentences. These also represent degradations introduced by packet switched networks (length micropause: 30 ms, 50 ms, 100 ms)
- EVS (Enhanced Voice Services, 3GPP codec) speech coding at bitrates between 13 and 48 kbit per second (FB and SWB)
- OPUS (IETF codec) speech coding at bitrates of 13 (SWB) and 24 kbit per second (FB)
- AMR (Adaptive Multi Rate, ETSI/3GPP codec) speech coding at bitrates of 13 (WB) and 24 kbit per second (WB)
- Low and medium levels of bandlimited pink noise, bandlimited between 7 and 24 kHz. This degradation checks the behavior of subjects in the upper wideband frequency range
- Low and medium levels of bandlimited pink noise, bandlimited between 14 and 24 kHz. This degradation checks the behavior of subjects in the upper super wideband frequency range

These degradations all still provide what is considered to be high quality speech in the classic telephony world, which uses a standard bandwidth of 3.5 kHz. The focus of this paper, however, is on the impact of small degradations in a high-quality Hi-Fi context. In this context the classic bandwidth is roughly the lowest quality that is acceptable, while pink noise degraded speech with an SNR of 30 dB can already be considered as severely degraded speech.

To be able to compare the speech quality of the different degradations with exactly the same speech material, each degradation is applied to all speech files. This gives 200 unique speech files in total: 25 conditions (24 degradations + 1 reference, see Table II) each applied to all eight speech files.

III. Subjective Test

A. Experimental Procedure

Different subjective test procedures can be used, such as double-stimulus or triple-stimulus methods described in ITU-R BS.1116 [30]. These two methods, however, focus on the assessment of small magnitudes of quality degradations. Therefore, they are not suited for assessing speech quality in the telephony

TABLE I
OPINION SCORES

Opinion Score	Opinion Score Absolute Category Rating Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

P.800 Absolute Category Rating (ACR) scale used in the subjective experiment [31]. If scores are averaged over a large set of subjects the resulting number is called a Mean Opinion Score (MOS), referred to as a MOS-LQS (Mean Opinion Score - Listening Quality Subjective).

context. In telephony the goal is not to provide a transparent link between the mouth of the talker and the ear of the listener, but to provide the best speech quality at the ear of the listener. To improve the end-to-end perceived speech quality, telephony links may therefore apply speech enhancement techniques. Two examples of speech quality enhancement are timbre optimization and noise suppression that also suppresses small amounts of breathing noise. To allow for possible improvements in the quality of the reference recording, only absolute quality assessments can be used. Therefore, in telephony context speech quality assessment, no explicit reference is provided. Hence, the subjective test procedure used in this paper largely follows the standard P.800 Absolute Category Rating (ACR) procedure used in the telecommunication industry [31].

A detailed overview of this procedure is given in Appendix II of ITU-T Recommendation P.863 [15]. In our subjective test 24 subjects were used in an Absolute Category Rating (ACR) experiment. The youngest subject was 19 and the oldest was 77. Each subject judged the quality of all 200 speech files using one out of eight different random orders. Subjects give each speech file a single opinion on the five-point ACR scale (see Table I). Each speech file consists of two short Dutch sentences including the silent interval between these sentences. The final results of the subjective test are expressed in terms of Mean Opinion Scores for Listening Quality Subjective (MOS-LQS). The predictions made with objective perceptual measurement methods on the other hand, are referred to as Mean Opinion Scores for Listening Quality Objective (MOS-LQO).

A well-known problem in P.800 tests is the effect of an improved intelligibility and quality when the same sentence is pronounced twice. This is especially relevant in low quality conditions where the MOS is below 2.0. Therefore these tests use very large sets of reference sentences where each degradation uses a different sentence so that each sentence is presented only once. However, the subjective test of this paper is focused on high quality speech and the use of many sentences, that are presented once, would mask small differences in speech quality. Furthermore, even in the worst degraded files intelligibility is high. To allow for detecting small quality differences in our ACR experiment we therefore use only eight sentences that are all degraded by all 25 experimental conditions.

All experiments are carried out in a silent room with a background noise level below 35 dB(A). A diotic (i.e., the same signal to both ears) presentation is used over high quality diffuse field equalized (up to 20 kHz) headphones (HEAD acoustics HPS IV amplifier/equalizer in combination with a Sennheiser HD600). The speech files are presented at a nominal Sound Pressure Level (SPL) of 73 dB(A) SPL.

B. Categorization of the Subjects into Poor, Good and Excellent Hearing

Before the experiment subjects are subdivided in one of three categories based on their hearing capabilities. This provides a deeper insight into the impact of the quality of the subjects hearing on the perceived speech quality. The first category has excellent hearing capabilities and are thus in general the younger subjects. The second category has good hearing, while the last category has degraded hearing. Degraded hearing can be caused either by age or by hearing loss. Hereafter these categories are referred to as excellent, good and poor hearing, respectively.

For categorization, subjects are presented with three band passed noise signals at a level of 53 dB SPL and are asked to carry out a small detection experiment. In this test, subjects indicated the audibility of a random on/off signal by hand raising. The first signal is 3.5–24 kHz band passed noise and is used as a sanity check to see if subjects have an acceptable minimum hearing threshold. Subjects are only accepted if they are able to detect this signal with a maximum amplification of 5 dB (58 dB SPL). This value is chosen in order to have an acceptable rejection rate for subjects older than 55 years. Using this procedure two subjects were rejected because they had too poor hearing capabilities.

Next, a 7–24 kHz band passed signal is used. If subjects are unable to detect this signal, they fall into the category poor hearing. Finally, a 14–24 kHz band passed signal is presented. Subjects that are not able to detect the noise in this signal fall into the category good hearing, while subjects that do detect the noise fall into the category excellent hearing. For a balanced experiment design, each category consists of eight subjects: four males and four females.

Note that with increasing age the quality of our ears drops significantly. Subjects over 65 years old will most likely always fall into the category poor hearing. In our test the youngest subject with poor hearing was 41, whereas the oldest subject in the category excellent hearing was 43. Furthermore, all subjects older than 63 fell in the category poor hearing or did not meet the minimum requirement of detecting the 3.5–24 kHz band passed noise signal.

C. Training of the Subjects

After having categorized the subjects into the three groups of hearing (poor, good and excellent) they are trained. First by presenting them with all eight reference files and the worst degradations for band limitation, noise, micropauses and codecs. These worst degradations correspond to conditions 4, 6, 12 and 21 in Table II.

TABLE II
Overview of the Subjective Results

CONDITION	MOS-LQS(CI95) per ear/subject group				F(2,23) ANOVA statistics
	all	poor	good	excellent	
01 FB 20-24000 Hz	4.35(.11)	4.19(.19)	4.58(.17)	4.30(.21)	
02 SWB 20-14000 Hz	4.32(.11)	4.20(.20)	4.53(.16)	4.23(.19)	
03 WB 20-7000 Hz	3.84(.13)	4.31(.18)	3.97(.18)	3.27(.21)	*9.86, p=.001,* η^2*=.484*
04 NB 20-3500 Hz	2.60(.15)	3.27(.29)	2.42(.17)	2.11(.21)	
05 white noise snr 40dB	2.65(.15)	3.28(.24)	2.41(.22)	2.22(.26)	
06 white noise snr 30dB	2.07(.14)	2.64(.23)	1.84(.20)	1.77(.25)	
07 mp low activity 30ms	3.99(.13)	4.02(.21)	4.20(.21)	3.92(.24)	
08 mp low activity 50ms	3.55(.14)	3.50(.23)	3.78(.25)	3.53(.23)	
09 mp low activity 100ms	2.80(.15)	2.92(.23)	2.70(.26)	2.70(.27)	
10 mp high activity 30ms	2.93(.14)	3.34(.24)	3.05(.22)	2.63(.23)	
11 mp high activity 50ms	2.61(.14)	3.08(.24)	2.50(.25)	2.44(.21)	
12 mp high activity 100ms	2.27(.14)	2.70(.25)	2.03(.23)	2.16(.22)	
13 EVS FB 48 kbps	4.26(.11)	4.08(.18)	4.64(.14)	4.03(.22)	*6.43, p=.007,* η^2*=.380*
14 EVS FB 32 kbps	4.02(.14)	4.08(.20)	4.44(.18)	3.58(.30)	
15 EVS FB 24.4 kbps	4.01(.14)	4.02(.22)	4.27(.21)	3.78(.25)	
16 EVS SWB 24.4 kbps	3.99(.14)	4.03(.22)	4.27(.21)	3.78(.27)	
17 EVS SWB 13.2 kpbs	3.82(.14)	3.95(.21)	4.19(.19)	3.42(.27)	
18 OPUS CBR FB 24.4kbps	3.90(.14)	4.00(.20)	4.34(.18)	3.39(.29)	
19 OPUS CBR SWB 13.2kbps	3.31(.13)	3.61(.24)	3.53(.17)	2.84(.19)	
20 AMR WB 23.85 kbps	3.21(.15)	3.88(.23)	3.30(.20)	2.45(.23)	*7.95, p=.003,* η^2*=.431*
21 AMR WB 12.65 kbps	3.15(.15)	3.73(.23)	3.11(.22)	2.59(.24)	*6.43, p=.007,* η^2*=.380*
22 14 kHz hpn snr 30dB	3.72(.20)	4.20(.18)	4.56(.17)	2.47(.38)	*12.4, p<.001,* η^2*=.542*
23 14 kHz hpn snr 20dB	3.27(.22)	4.06(.19)	4.48(.18)	1.28(.12)	*111.4, p<.001,* η^2*=.914*
24 7 kHz hpn snr 30dB	2.84(.19)	4.13(.20)	2.88(.26)	1.56(.14)	*24.7, p<.001,* η^2*=.701*
25 7 kHz hpn snr 20dB	2.34(.21)	4.06(.19)	1.92(.26)	1.06(.06)	*40.2, p<.001,* η^2*=.793*

Mean Opinion Scores for Listening Quality Subjective (MOS-LQS) results for the 25 speech processing conditions and their 95% Confidence Intervals for each of the three groups separately and overall. Averages are calculated over the eight different sentence pairs used in the subjective test. To unveil overall effects of the 25 conditions between the three groups of subjects: poor (age 41-77), good (age 32–62) and excellent (age 19–43) hearing, a one-way ANOVA was executed, including the average of the 8 MOS-LQS scores on all 25 conditions as dependent variables. We solely present the ANOVA F(2,23)-statistic and accompanying p and η2-values of indisputable differences with at least $p < .01$.
Legend: mp: micropauses and hpn: high pass noise.

In a second training phase ten test files, that include reference files and some of the worst degradations, are presented, and subjects are asked to score them using the full scoring range on the five point scale from Table I. If subjects only used three out of the five categories, they were asked to run the test a second time and use as much as possible the full five-point scale. The same training procedure was used in the development of PESQ [13], [14] and POLQA [16], [17].

After training, the subjects are presented with one of eight random orders of 200 test files in ten batches of twenty files. Each of the 200 test files is a unique random choice from 25 conditions (24 degradations and the fullband (FB) reference). Each condition is applied to eight different speech samples. These speech samples originate from four voices and each speech samples consists of two Dutch-spoken sentences with a short pause. After every batch of twenty files there is a small break. Between batch five and six there is a longer break in order to prevent subjects from getting tired. The total test duration, excluding breaks, is 30 minutes.

D. Subjective Test Results

Following the introduction on the design of the subjective test, the results are presented in three parts. These three parts relate to the three parts in the research question posed in the Introduction. In total, four distinct ANalysis Of VAriance (ANOVA) have been executed to determine whether or not there were differences between groups due to the study's parameters [32]. Each ANOVA is focused on investigating the impact of a different parameter of the research design, namely impact of: 1.a) the quality of hearing of the subjects, 1.b) audio bandwidth in relation to the quality of hearing, and 1.c) used coding techniques. The coding impact is investigated separately for two different bitrates, 24 and 13 kbps, giving four ANOVAs in total.

Additionally, one t-test is reported to test whether the most critical subject can hear the difference between SWB and FB. The statistical tests' degrees of freedom are calculated based on the number of groups and the total number of data samples. The study's design and the forthcoming analysis determine these calculations. For details on these calculations, we refer to [32].

The ANOVAs executed on the subjective MOS-LQS results allow the assessment of possible significant differences between the means of data subsets [32]. ANOVA results are presented in the form of its F-statistic, its two degrees of freedom, and the accompanying probability-value *p*, the probability that the null-hypothesis is true, and the proportion of variance accounted for, given by η2 (eta squared) or ηp2 (partial eta squared). For more background on statistical power analysis, we refer to Cohen's standard text on this topic. This work describes r2 and η2 in detail, among several others [33]. For an excellent handbook on applied statistics and a treatment of ANOVAs, we refer to [32].

The results of the first ANOVA and an overview of all MOS-LQS are given in Table II. This includes the F(2, 23)-statistic, p-values and η2-values. The first ANOVA focused on the differences between the three groups of hearing. The results follow from a one-way ANOVA, which investigates the *overall* effects of the 25 conditions between the three groups of hearing. The *average* MOS-LQS scores on all 25 conditions are taken as dependent variables. Here, we assess whether or not there is an *overall* difference between the three groups of hearing. The null-hypothesis is that there is no difference. This one-way ANOVA clearly shows indisputable differences between the three groups of hearing. In Table II only the significant results with a $p < 0.01$ are shown.

Fig. 1 gives the MOS-LQS results of the poor hearing group with respect to the excellent hearing group. This figure further illustrates the significance of the differences between both groups in their judgement on the 25 conditions.

The subjective MOS-LQS results on the standard narrowband speech signals show a rather low score of 2.6. In the POLQA standardization [15]–[17], standard listening tests were used and narrowband speech signals scored around 3.5 MOS-LQS. This difference is caused by the focus of the subjective test on the high-quality region, consequently leading to a warping of the MOS-LQS scale.

The second ANOVA is executed to unveil the impact of audio bandwidth in relation to the quality of hearing. The first four conditions in Table II give the MOS-LQS scores for poor, good and excellent hearing for the different bandwidths. This repeated measures ANOVA included all MOS-LQS scores with bandwidth (4 levels) and stimulus (8 levels) as within-subject factors and the three groups of hearing as between-subject factors. Here, we assess if there is a difference between the 4 bandwidths among the 3 groups of hearing. The null-hypothesis is that there is no difference between the bandwidths and the three groups of hearing. In line with expectations, this ANOVA unveiled a general difference among the four bandwidths, F(3,19) = 48.2, p < .001, ηp2 = .884. Subsequently, the three groups of subjects were compared with each other. No difference was found between FB and SWB. However, strong differences were found between SWB and WB (F(1, 21) = 31.2, p < .001, ηp2 = .598) and between WB and NB (F(1,21) = 91.4, p < .001, ηp2 = .813). Both results are in line with expectations.

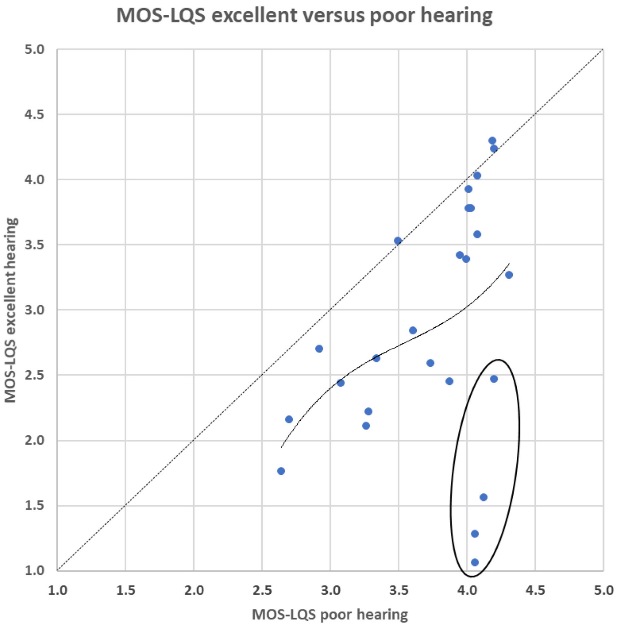

Fig. 1. Subjective MOS-LQS scores for the group with excellent hearing versus the subjective MOS-LQS scores for the group with poor hearing. Each point represents the averaged observed MOS-LQS for one of the 25 conditions for the poor hearing group versus the averaged observed MOS-LQS for the excellent hearing group. The line Y = X represents the expected regression when there would be no difference in behavior between the two groups. The third order polynomial fit gives the optimum mapping between the poor and excellent hearing groups, showing a strong bias in opinion between the groups. The most prominent significant differences, indicated by the ellipse around the data points, are for the high band passed noise degradations, conditions 22–25 in Table II.

An additional t-test was carried out to determine if the most critical subject can hear a difference between SWB and FB. One female subject aged 24, showed to have significantly better hearing than her 7 peers as determined via the group's average (t(199) = 17.1, p < .001, paired, 1-sided). However, even this very sensitive subject showed no difference in MOS-LQS scores between FB and SWB, with both scores being 3.63.

A third and fourth ANOVA are executed focused on the differences between codecs using the different bitrates for 24 kbps and 13 kbps, respectively. Both ANOVAs are repeated measures multivariate. Speech files are used as factors and the hearing group as between-subject factor. This was complemented with pairwise comparisons, including a Bonferroni adjustment for multiple comparisons.

The third ANOVA assessed whether there is a difference between the four 24 kbps CODECs, given by conditions 15, 16, 18 and 20 in Table II. The null-hypothesis is that there is no difference. This ANOVA showed however, that the MOS-LQS scores for the four 24 kbps CODEC conditions differed significantly, F(3,19) = 20.2, p < .001, ηp2 = .761. This is explained by the AMR WB that scored significantly lower than the other three 24–kbps conditions: EVS FB (p < .001), EVS SWB (p < .001), and OPUS (p < .001).

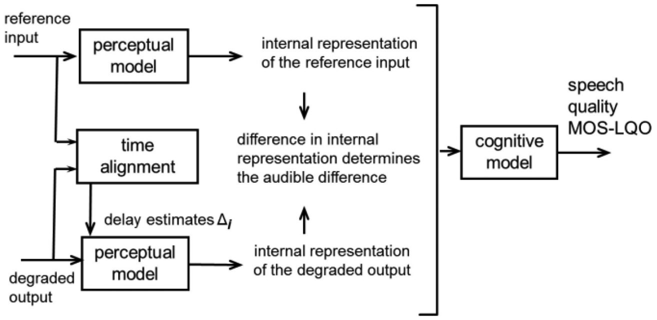
reference input
perceptual model
internal representation of the reference input
time alignment
difference in internal representation determines the audible difference
delay estimates Δ_I
perceptual model
internal representation of the degraded output
degraded output
cognitive model
speech quality MOS-LQO

Fig. 2. Overview of the basic strategy used in a perceptual measurement approach. A computer model of the subject, which consists of a perceptual and a cognitive model, is used to compare the degraded output with the reference input. Alignment information is used in the perceptual model, after being derived from the time signals in the time alignment module. The objective MOS score is referred to as MOS-LQO (Mean Opinion Score Listening Quality Objective).

The fourth ANOVA assessed if there is a difference between the three 13 kbps CODECs, given by conditions 17, 19 and 21 in Table II. The null-hypothesis is that there is no difference. However, this ANOVA unveiled that the MOS scores on the three 13 kbps CODEC conditions also differed significantly, F(2, 20) = 11.6, p < .001, ηp2 = .536. This is explained by the EVS SWB that scores significantly higher than both AMR WB (p < .001) and OPUS CBR SWB (p = .002), as can be seen in Table II.

IV. Objective Quality Assessment

A. Objective Quality Assessment Approach

The most widely used objective speech quality assessment tool is POLQA as standardized by the ITU-T as recommendation P.863 (Perceptual Objective Listening Quality Prediction [15]–[17]). POLQA constructs internal representations of the reference and degraded speech signal based on a perceptual model. A time-frequency decomposition in terms of a windowed frequency spectrum is used. The physical representation is given by the power spectra in dB per Hertz per time window. This representation is mapped to an internal psycho physical representation given by the loudness spectra in Sones per Bark per time window. The difference function is interpreted by a cognitive model and represents a loudness spectrum in Sones per Bark per time window. The output of this cognitive model is used to predict the quality of the speech file on a five-point MOS scale. The predicted MOS values are denoted as MOS-LQO (Mean Opinion Score Listening Quality Objective). A schematic overview of this objective measurement procedure is given in Fig. 2.

In the mapping from physical to psycho physical representation the average absolute hearing threshold is used. The current standard uses a threshold averaged for subjects over the age range 16 to 70 years. For a correct prediction of the speech quality, a correct threshold is of importance. Especially for high quality speech coding with a bandwidth above 7 kHz, small difference in this threshold can have a large impact on the predicted speech quality.

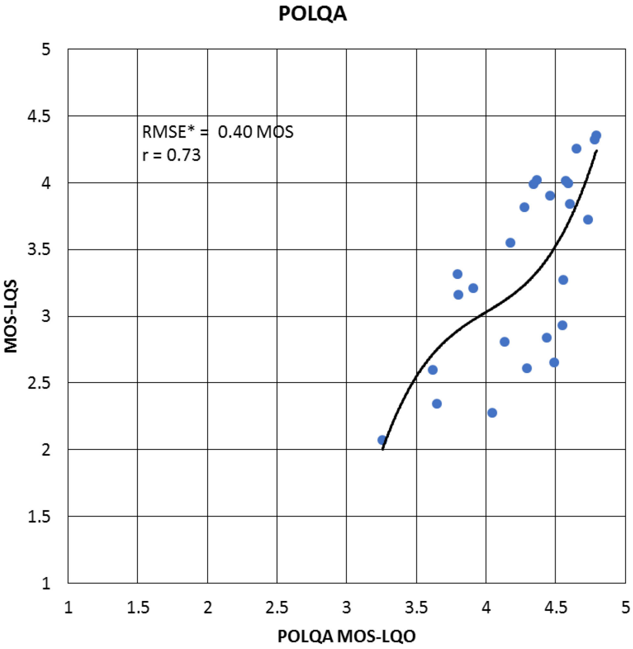

Fig. 3. Subjective (MOS-LQS) versus objective (MOS-LQO) results obtained with ITU-T Rec. P.863 POLQA on the per condition. Each point represents the observed MOS-LQS for one of the 25 conditions, using the results for all three hearing groups, versus the objectively predicted MOS (MOS-LQO). The line represents the optimal third order polynomial fit, using a standard regression approach (see [15], [17]). The average bias of around 0.8 MOS between subjective and objective data is caused by the experimental context.

POLQA uses a standardized method for assessing the accuracy of the predicted MOS-LQO. This standardized method uses a modified Root Mean Squared Error (RMSE) that considers the reliability of each subjectively obtained MOS-LQS [34]. This modified RMSE is obtained by subtracting the 95% confidence interval from the measured deviation between the MOS-LQS and MOS-LQO values before calculating the squared error and is called the RMSE*. For each assessed database, the RMSE* is calculated using a 3rd order monotonically increasing regression function. Thereby taking the context of the subjective experiment into account.

B. Objective Assessment Results Using ITU-T P.863 POLQA

The degraded speech signals used in the subjective experiment are evaluated using POLQA (ITU-T P.863 [15] in super wideband mode). This allows to test if POLQA is capable of predicting MOS scores for high quality voice links. The result of this evaluation is given in Fig. 3 for the per condition averaged MOS and shows a strong bias in the optimal regression curve: all data points are below the theoretical optimal regression curve MOS-LQS = MOS-LQO. The average bias is around 0.8 MOS and the lowest MOS-LQS of around 2.0 is mapped to a MOS-LQO of around 3.5. The bias is expected, as the lowest speech quality in the experiment is much higher than the lowest speech quality used in standard P.800 subjective testing. In our experiment the standard narrowband telephony quality, using a low pass filtering of 3.5 kHz, obtains a MOS-LQS of 2.6.

The same condition in the context of a standard subjective experiment with a wide range of degradations, shows a MOS-LQS of about 3.5. This effect is known as the "range equalizing bias" effect [35].

The results in Fig. 3 show a rather large RMSE* of 0.40 MOS and low correlation of 0.73. These values are worse than normally obtained when using POLQA to predict subjective scores. In the development of the POLQA measurement standard an average RMSE* of 0.20 MOS and an average correlation of 0.93 were found over a set of 25 (super)wideband databases [17]. The worst database in this set showed an RMSE* of 0.23 MOS with a correlation of 0.91. In a validation of POLQA on three databases containing unknown degradations, the worst RMSE* was 0.24 MOS with a correlation of 0.87 [17]. All these performance numbers are significantly better than obtained with the high-quality voice experiment of this paper. A major reason for the higher RMSE* and lower correlation is the limited range of MOS-LQO. While standard subjective tests generally have a range of MOS-LQO between 1.5 and 4.7, our dataset spanned a range of 3.5 to 4.7 MOS-LQO. This limited range in MOS-LQO gives an expansion of the subjective MOS-LQS values, which are consequently more difficult to predict.

A subset analysis is made using only the codec and bandwidth limitation conditions to further investigate the performance of POLQA. These conditions represent the most widely used conditions in the assessment of high-quality voice links. In this subset the artificial noise conditions and micropause results are omitted, as these conditions are challenging to predict. The results are given in Fig. 4 and show an extremely good performance of POLQA: with 0.05 MOS the RMSE* is extremely low and the correlation is extremely high with 0.98.

Over the period 2015-2018 a number of improvements were implemented for POLQA of which the most important ones are:

- Decreased 'shift-jitter' by using a higher FFT overlap, 0.75 instead of 0.5.
- Decreased complexity by using a single run in the calculation of the disturbances omitting the second run used for a separate calculation of loud degradations
- Extension towards full audio bandwidth (>20 kHz)
- Improved modelling of the impact of modern speech codecs
- Adequate consideration of extended micropauses in the speech signal as caused by time-variant transmission in VoIP or general packet-switched connections

A number of these improvements were developed on a limited set of data. Especially, the introduction of micropauses and the extension towards fullband could only be trained on a limited set of data.

The latest POLQA version from March 2018 [15] was optimized with a fullband correction parameter and a re-optimization of the micropause parameter, based on the subjective results of this paper. The introduction of micropauses is modeled in the 2018 version by a special detector that models the impact of the introduction of micropauses when the length of the micropause exceeds 40 ms. The perceptual model is expected to correctly quantify the impact of micropause below this threshold. This test showed that this lower bound had to be lowered to 28 ms. This was changed in the latest development version of POLQA and was trained on all available data including the data of this experiment. No significant performance drop on any of the 25 POLQA databases available in ITU supplement 23 [36] was allowed. The results of the latest development version of POLQA are given in Fig. 5 and 6, for all subjective data and for subset of codecs and bandwidth limitation conditions, respectively. A significant better performance is found when compared to the current standard. The source code differences between the current POLQA [15] and the latest development version are marginal and consist only of an extra fullband degradation indicator and the retraining of a small set of perceptual modelling parameters, including the micropause parameters.

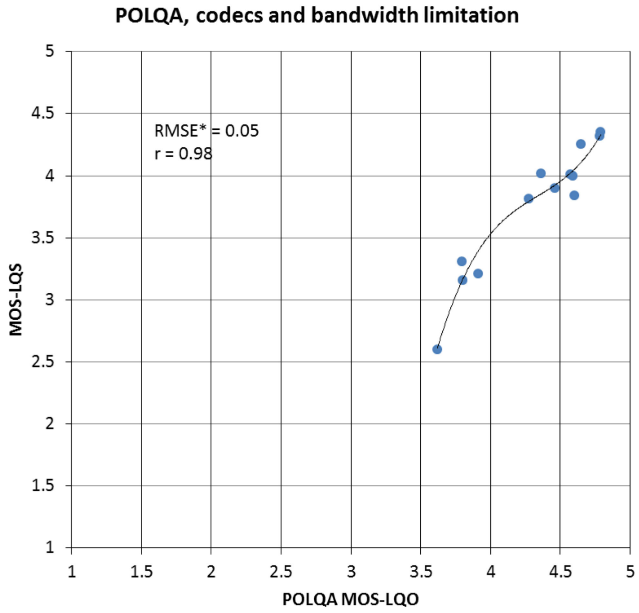

Fig. 4. Subjective (MOS-LQS) versus objective (MOS-LQO) results obtained with ITU-T Rec. P.863 POLQA on the per condition data for the codec/bandwidth subset. Each point represents the observed MOS-LQS for one of the 13 conditions, using the results for all three hearing groups, versus the objectively predicted MOS (MOS-LQO). The line represents the optimal third order polynomial fit, using a standard regression approach (see [15], [17]). The bias of around 0.8 MOS between subjective and objective data is caused by the experimental context.

The objective assessment results are only valid for the average score over all subjects. An imbalance over the poor, good and excellent hearing groups, may result in a significant higher RMSE* and lower correlation.

V. Discussion

The results of this paper are twofold. On the one hand, we have the results of the subjective speech quality experiment, using high quality fullband speech recordings (20 Hz – 24 kHz). On the other hand, we have the results from the objective perceptual measurement analysis, using the standard perceptual measurement method ITU-T Rec P.863 POLQA [15]–[17].

The focus of the subjective experiment was on the impact of degradations found in modern fullband speech coding schemes. This impact can depend significantly on the quality of subject's

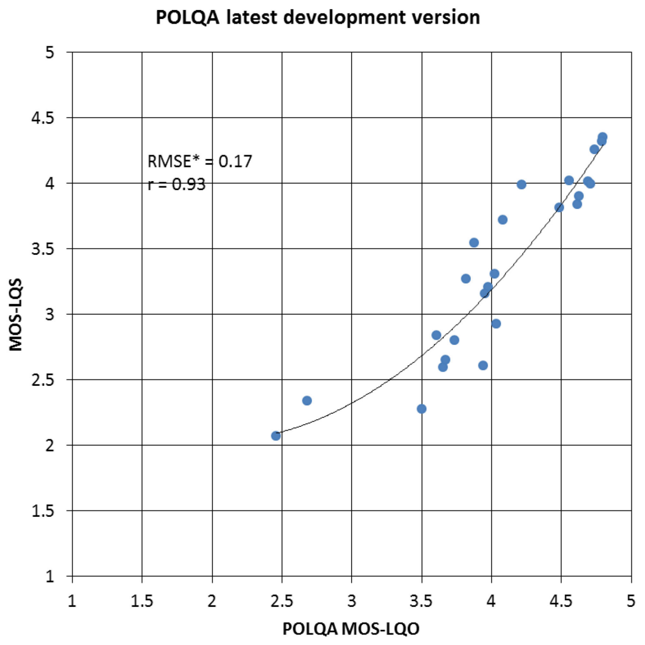

Fig. 5. Subjective (MOS-LQS) versus objective (MOS-LQO) results obtained with the latest development version of ITU-T Rec. P.863 POLQA on the per condition data. Each point represents the observed MOS-LQS for one of the 25 conditions, using the results for all three hearing groups, versus the objectively predicted MOS (MOS-LQO). The line represents the optimal third order polynomial fit, using a standard regression approach (see [15], [17]). The bias of around 0.8 MOS between subjective and objective data is caused by the experimental context.

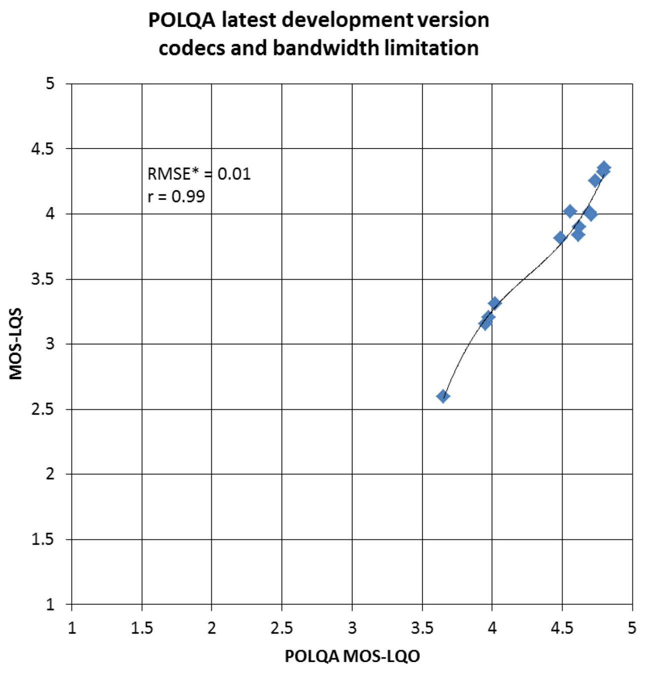

Fig. 6. Subjective (MOS-LQS) versus objective (MOS-LQO) results obtained with the latest development version of ITU-T Rec. P.863 POLQA on the per condition data for the codec/bandwidth subset.. Each point represents the observed MOS-LQS for one of the 13 conditions, using the results for all three hearing groups, versus the objectively predicted MOS (MOS-LQO). The line represents the optimal third order polynomial fit, using a standard regression approach (see [15], [17]). The bias of around 0.8 MOS between subjective and objective data is caused by the experimental context.

hearing. Therefore, our subjects were assessed on the quality of their hearing and, subsequently, divided in three groups: poor, good, and excellent hearing. As expected, the analysis of the subjective data unveiled a large difference between the three groups. Moreover, the results showed that even for the excellent hearing group there is no significant difference in perceived speech quality between a clean fullband reference speech signal and a super wideband speech signal, with frequency components up to 24 kHz and 14 kHz, respectively. A separate t-test showed that this result even holds for the most critical subject in the test. We conclude that to obtain high speech quality, a fullband speech codec's input and output signal should be bandlimited to 14 kHz.

The subjective experiment also unveiled that subjects with poor hearing, most likely all subjects older than 65 years, do not perceive an improvement in speech quality when the bandwidth is extended from wideband (7 kHz) to super wideband (14 kHz). This raises the question whether or not one should adapt the coding strategy towards specific user groups, for instance based on age. Especially for very low bitrates, the optimal coding strategy depends on the capability of the user to detect degradations in the upper frequency bands. For example, elderly users may benefit more from accurate coding of the frequency range between 200 and 4000 Hz than from an extended bandwidth. The subjective results on the EVS, OPUS and AMR codecs, showed that EVS has the best subjectively perceived quality, while AMR has the lowest subjectively perceived quality. Both results hold for 24 kbps and 13 kbps.

The objective results are obtained by ITU-T Rec P.863 POLQA [15], the worldwide accepted perceptual measurement standard for assessing speech quality. These objective results show that POLQA is capable of predicting degradations, as introduced by modern speech codecs including effects of bandwidth limitation. However, for assessing high quality voice links in general, the performance of the current ITU standard is too low for accurate speech quality predictions. Both for micropauses, as often found with time-variant VoIP (packet-switched) connections, and for severe, audible, degradations above 7 kHz, the performance is too low.

The subjective results showed that differences in perceived speech quality between subjects with excellent hearing and poor hearing can be very large. Consequently, when separate predictions are made for these groups, slightly different perceptual models should be used. For the average hearing, a modified version of ITU-T P.863 POLQA was constructed, which shows a significant improvement over the current standard in overall performance on our high-quality speech dataset. This latest improved POLQA development version has an extra fullband indicator and has a small set of model parameters retrained. The correlation between objective and subjective results increases from 0.73 for the current standard to 0.93 for the latest improved development version. The Root Mean Squared Error (RSME) from which the 95% confidence interval is subtracted, the so called RMSE*, drops from 0.38 MOS to 0.17 MOS for this latest development version. Comparison between subjective results and objective POLQA predictions show that codec and bandwidth limitation conditions are predicted with high accuracy. For both the current POLQA standard and the latest improved development version, the correlation results for this subset are excellent: above 0.98 for both versions, while the RMSE* is extremely low: below 0.05 MOS.

VI. Conclusion

A major conclusion based on the carried out ITU-T P.800 absolute category rating experiment, is that there is no significant difference between a clean fullband speech reference file with frequency components up to 24 kHz and a super wideband representation that is bandlimited to 14 kHz. This result even holds for the most critical subjects that are clearly able to detect frequencies above 14 kHz.

Furthermore, the results show that the differences in perceived speech quality between subjects with excellent hearing and poor hearing can be very large. Therefore, slightly different perceptual models should be used for both groups. For the average hearing, a modified version of ITU-T P.863 POLQA is constructed, which shows a significant improvement in overall performance over the current standard: The RMSE* drops from 0.38 MOS for the current standard to 0.17 MOS for the latest improved development version. Comparison between subjective results and objective POLQA predictions showed that codec and bandwidth limitation conditions are predicted with high accuracy: a RMSE* of below 0.05 MOS.

References

[1] R. V. Cox, S. F. de Campos Neto, C. Lamblin, and M. H. Sherif, "ITU-T coders for wideband, super wideband, and fullband speech communication [Series Editorial]," *IEEE Commun. Mag.*, vol. 47, no. 10, pp. 106–109, Oct. 2009.

[2] A. Ramo and H. Toukomaa, "Subjective quality evaluation of the 3GPP EVS codec," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, 2015, pp. 5157–5161.

[3] M. Dietz *et al.*, "Overview of the EVS codec architecture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, 2015, pp. 5698–5702.

[4] K. Vos, K. V. Sørensen, S. S. Jensen, and J.-M. Valin, "Voice coding with opus," in *Proc. 135th Conv. Audio Eng. Soc.*, Oct. 2013, pp. 722–731.

[5] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2085–2095, Oct. 2013.

[6] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers Psychol.*, vol. 5, Jun. 2014, Art. no. 587, doi: 10.3389/fpsyg.2014.00587.

[7] C. H. Shadle and S. J. Mair, "Quantifying spectral characteristics of fricatives," in *Proc. IEEE Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, USA, 1996, pp. 1521–1524.

[8] K. Shoji, E. Regenbogen, J. D. Yu, and S. M. Blaugrund, "High frequency components of normal voice," *J. Voice*, vol. 5, pp. 29–35, 1991.

[9] B. C. J. Moore and C. T. Tan, "Perceived naturalness of spectrally distorted speech and music," *J. Acoust. Soc. Amer.*, vol. 114, pp. 408–419, 2003, doi: 10.1121/1.1577552.

[10] H. Fletcher and R. H. Halt, "The perception of speech and its relation to telephony," *J. Acoust. Soc. Amer.*, vol. 22, no. 2, Mar. 1950, Art. no. 89.

[11] M. Cernak, A. Asaei, and A. Hyafil, "Cognitive speech coding: Examining the impact of cognitive speech processing on speech compression," *IEEE Signal Process.*, vol. 35, no. 3, pp. 97–109, May 2018.

[12] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, ITU-T Rec. P.862.2, International Telecommunication Union, Geneva, Switzerland, Nov. 2007.

[13] 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, Oct. 2002.

[14] 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, Oct. 2002.

[15] *Perceptual Objective Listening Quality Prediction*, ITU-T Rec. P.863, International Telecommunication Union, Geneva, Switzerland, Mar. 2018.

[16] J. G. Beerends *et al.*, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement, part I—Temporal alignment," *J. Audio Eng. Soc.*, vol. 61, pp. 366–384, Jun. 2013.

[17] J. G. Beerends *et al.*, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement, part II—Perceptual model," *J. Audio Eng. Soc.*, vol. 61, pp. 385–402, Jun. 2013.

[18] L. A. Ekman, V. Grancharov, and W. B. Kleijn, "Double-ended quality assessment system for super-wideband speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 558–569, Mar. 2011.

[19] L. O. Nunes, L. W. P. Biscainho, B. Lee, A. Said, T. Kalker, and R. W. Schafer, "Degradation type classifier for full band speech contaminated with echo, broadband noise, and reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2516–2526, Nov. 2011.

[20] J. Abel, M. Kaniewska, C. Guillaume, W. Tirry, and T. Fingscheidt, "An instrumental quality measure for artificially bandwidth-extended speech signals," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 384–396, Feb. 2017.

[21] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: An objective speech quality model," *EURASIP J. Audio, Speech Music Process.*, vol. 2015, 2015, Art. no. 13.

[22] C. Sloan, N. Harte, D. Kelly, A. Kokaram, and A. Hines, "Objective assessment of perceptual audio quality using ViSQOLAudio," *IEEE Trans. Broadcast.*, vol. 63, no. 4, pp. 693–6705, Dec. 2017.

[23] P. Počta and J. G. Beerends, "Subjective and objective assessment of the listening quality of customer support waiting loops," *Acta Acustica United Acustica*, vol. 105, pp. 392–400, 2019.

[24] F. B. Gelderblom, T. V. Tronstad, and E. M. Viggen, "Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 583–594, Mar. 2019.

[25] S. Islam, M. Rashid, and M. Tarique, "Performance analysis of WiMax/WiFi system under different codecs," *Int. J. Comput. Appl.*, vol. 18, no. 6, pp. 13–19, Mar. 2011.

[26] M. A. Qureshi, A. Younus, M. Saeed, F. A. Sidiqui, N. Touheed, and M. S. Qureshi, "Comparative study of VoIP over WiMAX and Wi-Fi," *Int. J. Comput. Sci. Issues*, vol. 8, no. 3, pp. 433–437, May 2011.

[27] M. Alahmadi, Y. Cinar, H. Melvin, and P. Pocta, "Investigating the extent and impact of time-scaling in WebRTC voice over IP traffic under light, moderate and heavily congested Wi-Fi APs," in *Proc. 5th ISCA/DEGA Workshop Perceptual Qual. Syst.*, 2016, pp. 87–91, doi: 10.21437/PQS.2016-19.

[28] *Objective Measurement of Active Speech Level*, ITU-T Rec. P.56, International Telecommunication Union, Geneva, Switzerland, Mar. 1993.

[29] J. G. Beerends and I. Beerends, "On the assessment of high-quality voice recordings including voice post processing," *J. Audio Eng. Soc.*, vol. 63, pp. 174–183, Mar. 2015.

[30] *Methods for the Subjective Assessment of Small Impairments in Audio Systems*, ITU-R Rec. BS.1116, International Telecommunication Union, Geneva, Switzerland, Feb. 2015.

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

[32] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 7th ed. Hoboken, NJ, USA: Wiley, 2018.

[33] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Mahwah, NJ, USA: Lawrence Erlbaum, 1988.

[34] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, ITU-T Rec.1401, International Telecommunication Union, Geneva, Switzerland, 2012.

[35] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests—A review," *J. Audio Eng. Soc.*, vol. 56, pp. 427–451, Jun. 2008.

[36] *ITU-T Coded-Speech Database*, ITU-T Suppl. 23 to P.863, International Telecommunication Union, Geneva, Switzerland, 1998.

John G. Beerends received the B.Eng. degree in electrical engineering from the Polytechnic Institute of The Hague, Netherlands, in 1975, the M.Sc. degree in physics from the University of Leiden, Leiden, Netherlands, in 1984, and the Ph.D. degree in psychoacoustics from the Technical University of Eindhoven, Eindhoven, Netherlands, in 1989, working on pitch perception and resulted in a Philips patent on a pitch meter. In 1983, he was awarded a prize of DFl 45000 by Job Creation for the further development of his patented asymmetric loudspeaker enclosure. In 1989, he joined KPN Research where he worked on audio and video quality assessment, audio-visual interaction, and on speech/audio coding leading to several patents and two perceptual measurement methods for assessing audio quality. The first one dealt with telephone-band speech and was standardized in 1996 as ITU-T Rec. P.861 (Perceptual Speech Quality Measure, PSQM), the second one with wideband audio and was integrated into ITU-R Rec. BS.1387 (1998, Perceptual Evaluation of Audio Quality, PEAQ). From 1996 to 2002, he worked on the objective measurement of the quality of video and speech. The work on speech quality, partly carried out with researchers from British Telecom, was focused on improving PSQM and was standardized in 2001 as ITU-T Rec. P.862 (Perceptual Evaluation of Speech Quality, PESQ). The work on video quality led to several patents and a measurement method for objective, perceptual, assessment of video quality, standardized in 2008 by the ITU-T as Rec. J.247 (Perceptual Evaluation of Video Quality, PEVQ). In 2003, he joined The Netherlands Organization for Applied Scientific Research (TNO), where he worked on the objective measurement of speech intelligibility, fullband speech quality, degradation decomposition, hearing aid quality, video quality. During 2003–2010, he worked on the development of the follow up of PESQ P.862 in a joint effort with OPTICOM and SwissQual. This work resulted in ITU-T Rec. P.863 (POLQA) in 2011. He is currently working on extending the perceptual measurement approach toward intelligibility and acoustic domain measurements (loudspeaker reproduction quality, including the impact of the reproduction room) and on the glass box modeling of audio, speech, video, and data services. He has authored more than 100 (conference) papers/ITU contributions and 35 patents. Dr. Beerends was the recipient of an AES Fellowship Award for his work on audio and video quality measurement in 2003.

Niels M. P. Neumann received the M.Sc. degree in mathematics and physics from Radboud University Nijmegen, Nijmegen, The Netherlands, in 2016. After graduation, he joined TNO, The Netherlands Organization for Applied Scientific Research, working on a wide range of topics, such as predicting railway degradations and applications of quantum computers. He has worked on models for the objective measurement of speech quality and contributed to maintaining ITU-T Rec. P.863 (POLQA).

Egon L. van den Broek received the M.Sc. degree in artificial intelligence (AI) from Radboud University Nijmegen (RU), Nijmegen, The Netherlands, in 2001, the Ph.D. degree in content-based image retrieval (CBIR) from RU in 2005, and the Ph.D. degree in affective signal processing (ASP) from the University of Twente (UT), Enschede, The Netherlands, 2011. Since 2004, he has been a (Part-Time) Assistant Professor. As such, he has coordinated, developed, and lectured several courses and M.Sc. tracks and programs and guided more than 85 B.Sc., M.Sc., and Ph.D. students. Moreover, he is the Vice Chair of his faculty's Ethical Review Committee, Utrecht University, The Netherlands, and the Science Lead with the Center for Research on Data-Driven User eXperience (CRUX), The Netherlands. Additionally, he is a consultant (e.g., for TNO, Philips, and the United Nations) and an external expert for various agencies (e.g., of the European Union), as PC member of conferences, and on boards of advice. He has authored or coauthored more than 200 scientific articles and has four patents. He has been an invited and keynote speaker at both conferences and in the media. Dr. van den Broek was the recipient of several awards. He is the Editor-in-Chief for the *Open Computer Science* journal, Associate Editor for the journals *Pattern Recognition Letters* and *Frontiers in Human Media Interaction* as well as the editorial board member of several other journals.

Anna Llagostera Casanovas received the M.S. degree in telecommunication engineering and a European Masters in Language and Speech from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2006, and the Ph.D. degree in signal and image processing from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2011. She was a Postdoctoral Researcher with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. In 2012, she joined SwissQual AG, Switzerland, now integral part of Rohde & Schwarz, as a Research Engineer, where she works on the development of algorithms that assess the quality of audio and video signals. Her research interests include image, speech, video, and multi-modal signal representation and processing, source separation, and computer vision with applications to audio-visual data analysis. Dr. Llagostera Casanovas was the recipient of a Prospective Researcher Fellowship from the Swiss National Science Foundation in 2011.

Jovana Torres Menendez studied electronic engineering at the University of of Belgrade, Belgrade, Serbia, and multimedia and communications engineering at Friedrich-Alexander Universität, Erlangen, Germany. Her master's dissertation was on the validation and development of distortion dimension indicators for noisiness, loudness, and continuity for super-wideband speech signals. In 2016, she joined OPTICOM as a Software Developer.

Christian Schmidmer studied electronic engineering and received the M.S. degree (Diplom) from the University of Erlangen, Erlangen, Germany. He spent five years as a Scientist with the Audio Department, Fraunhofer Institute for Integrated Circuits, Erlangen, Germany (the home of mp3), mostly dedicated to the research of psychoacoustics and the development of perceptual measurement tools as well as audio codecs, contributing to the development of mp3. In 1997, he joined OPTICOM as CTO and Co-Owner. OPTICOM's core business is the development and IPR management for voice, audio, and video quality measurement algorithms. He is active in standardization bodies, such as ITU, VQEG, and ETSI. He has authored many scientific publications and frequently presents papers at conferences and workshops. He is one of the main developers behind the recommendations ITU-R BS.1387 (Perceptual Evaluation of Audio Quality, PEAQ), ITU-T P.563 (no-reference voice quality assessment, 3SQM), and ITU-T P.863 (POLQA).

Jens Berger received the Ph.D. degree in electrical engineering in the area of network- and system-theory from the Technical University of Kiel, Kiel, Germany, in 1998. He started his carrier with the Research Institute of Deutsche Telekom, Berlin, Germany. Since 2003, he has been with SwissQual AG, Switzerland, now integral part of Rohde & Schwarz, heading the Applied Research Department as Member of the Senior Management Team. Rohde & Schwarz SwissQual AG is one of the leading suppliers of measurement equipment in the segment of mobile network testing. For the past seventeen years, has been leading the working group: "Perceptual-based objective methods for voice, audio and visual quality measurements in telecommunication services" as Rapporteur in ITU-T SG12. He is further active in ETSI and other international standardization bodies. His work has contributed to several ITU-T standards for voice and video quality prediction.