http://dx.doi.org/10.17743/jaes.2016.0034

Quantifying Sound Quality in Loudspeaker Reproduction*

JOHN G. BEERENDS[1], *AES Fellow*, **KEVIN VAN NIEUWENHUIZEN**[2], **AND EGON L. VAN DEN BROEK**[2]
(john.beerends@tno.nl) (kevinvannieuwenhuizen@gmail.com) (vandenbroek@acm.org)

[1] *TNO, P.O. Box 96800, NL-2509 JE The Hague, The Netherlands*
[2] *Utrecht University, P.O. Box 80.089, NL-3508 TB Utrecht, The Netherlands*

We present PREQUEL: Perceptual Reproduction Quality Evaluation for Loudspeakers. Instead of quantifying the loudspeaker system itself, PREQUEL quantifies the overall loudspeakers' perceived sound quality by assessing their acoustic output using a set of music signals. This approach introduces a major problem: subjects cannot be provided with an acoustic reference signal and their judgement is based on an unknown, internal, reference. However, an objective perceptual assessment algorithm needs a reference signal in order to be able to predict the perceived sound quality. In this paper, these reference signals are created by making binaural recordings with a head and torso simulator, using the best quality loudspeakers, in the ideal listening spot in the best quality listening environment. The reproduced reference signal with the highest subjective quality is compared to the acoustic degraded loudspeaker output. PREQUEL is developed and, subsequently, validated, using three databases that contain binaurally recorded music fragments played over low to high quality loudspeakers in low to high quality listening rooms. The model shows a high average correlation (0.85) between objective and subjective measurements. PREQUEL thus allows prediction of the subjectively perceived sound quality of loudspeakers taking into account the influence of the listening room and the listening position.

0 INTRODUCTION

Over the past decades, models for the perceptual evaluation of audio signals have been introduced for a wide range of application areas. They allow assessment of the quality of time variant, nonlinear systems. As such, they are essential for quality assessment of low bit rate speech and audio coding, as used in the telecommunication [1, 2, 3, 4, 5] and music [6, 7, 8, 9] industries. These models assess the system's signal adaptive properties by feeding them real world speech and/or music signals. They measure the quality of the system's output signals by processing both a reference and a degraded signal, using a psychoacoustic model. The difference between the internal representations of both signals is assessed by a cognitive model, which provides an objective quality rating of the degraded signal (see Figure 1). This quality is expressed in terms of subjects' subjective Mean Opinion Score (sMOS), on a scale from 1 to 5 [10] (see Table 1). Please note that this approach quantifies the quality of the system's output and does not characterize it directly by a set of technical parameters such as frequency response, harmonic distortion, etc.

Table 1. Absolute category rating listening quality opinion scale [10]. The mean calculated over a set of subjects is called the subjective Mean Opinion Score (sMOS).

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

Traditional research in loudspeaker reproduction quality follows a classical approach. Instead of characterizing the perceived sound quality produced by the loudspeakers, one quantifies the loudspeaker system directly on the basis of a set of technical measurements. An extensive overview of this approach is given by Toole [11, 12, 13, 14]. Although a loudspeaker is to a large extent a linear time invariant system, quality assessment is difficult. This is due to the fact that a one dimensional input music signal (i.e., an ampli-

*Tel: +31-6-5161-2563; e-mail: john.beerends@tno.nl

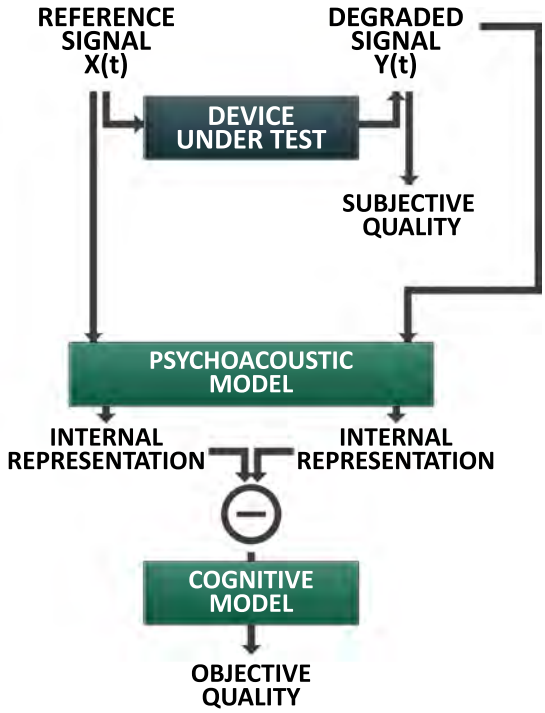

Fig. 1. Overview of the basic principle used in models for the perceptual evaluation of audio signals. The psychoacoustic and cognitive model are used to create an objective quality measure that uses the reference signal $X(t)$ and the degraded output, $Y(t)$, of the device under test. This objective quality measure is used to predict the subjective quality of the degraded signal in terms of a mean opinion score.

tude as a function of time) produces a four dimensional output (i.e., a pressure as a function of space and time). This output is only assessed using two pressure waves at the entrance of our ears, as a function of time. Furthermore, most music is enjoyed through the use of two or more loudspeakers that interact with the room and with each other. Room reflections, resonances and comb filtering have a major impact on the reproduction quality [14].

When placed in an anechoic room, loudspeaker reproduction quality is unsatisfying. In stereo, we get comb filtering between the outputs of the two loudspeakers and there is no sense of envelopment due the lack of lateral reflections. This is illustrated in Figure 9.7 of Toole's book [14], where measurements are shown for two loudspeakers located in an anechoic and a standard listening room. In a normal listening room, the comb filter effects between the two loudspeakers are reduced due to room reflections; however, they are still visible and audible when listening to a phantom center signal produced by left/right correlated noise or music. Furthermore room reflections themselves introduce comb filter effects.

Characterization of loudspeakers' radiation pattern into space has only limited value because the interaction between the listening room and the loudspeaker will determine the pressure waves at the entrance of our ears. When a loudspeaker is reproducing a music signal in a room, radiation pattern measurements —like on- and off-axis pressure response, power response and directivity index —can at best only predict the relative average perceived sound quality of the loudspeaker, even when it is supplemented with other technical measurements such as harmonic distortion and rub/buzz characterizations [12, 13, 15, 16]. A certain loudspeaker may sound excellent with one type of music signal, while the same setup may show low quality with another type of music signal. A high quality loudspeaker evaluated at a non-optimal listening spot in a room may sound worse than a low quality loudspeaker evaluated at the optimal listening spot. Although some approaches use signal properties at the listening place [17] or room characterizations [18, 19], no method exists yet that allows to assess the quality of a single musical fragment when reproduced over a loudspeaker set up in a certain room on a certain listening spot without having to measure any technical parameter of either the loudspeaker or the listening room.

Instead of quantifying the loudspeaker system with technical measurements, an unconventional approach can be chosen: quantification of the loudspeakers' perceived sound quality in normal listening rooms, with music signals, on the basis of a perceptual model that operates on binaurally recorded signals. Tan et al. [20, 21] described a model for predicting the effect of various forms of nonlinear distortions generated by electro-acoustic transducers on the perceived quality of speech and music signals. However, their subjective experiments made use of headphones to judge the audibility of loudspeakers' distortions. Hence, they did not take into account the influence of the listening room and listening position, which both have a significant impact on the listener's perceived loudspeaker reproduction quality. Gabrielsson et al. [22, 23] assessed the sound quality of loudspeakers directly using subjective experiments, by applying a decomposition of the acoustic output into perceptual dimensions (e.g., clearness, loudness, nearness, and spaciousness). However, they did not develop an objective measurement method using this data. Conetta et al. [24, 25] used the ideas of source localization, envelopment, coverage angle, ensemble width, and spaciousness to describe a model that successfully assesses the spatial audio perception quality. While this model is successful in its specified domain, it does not generalize to loudspeakers' overall perceived sound quality and is limited to a small number of high quality loudspeakers and listening environments.

Whereas previous research focused on the quantification of the loudspeaker system itself or on specific aspects of the acoustic output of loudspeakers, the aim of this paper is to generalize to loudspeakers overall perceived sound quality in a wide variety of environments, using a diverse set of stereo recorded music fragments that are played over the loudspeaker systems. This approach thus does not need any technical measurements on the loudspeaker, it only uses recordings of musical fragments played over the loudspeakers. Consequently this paper does not measure nor specifies any technical parameter for the loudspeakers that are used. This new perceptual modeling approach allows a direct comparison between the perceived quality of an excellent loudspeaker in a bad reproduction room/non-

 http://dx.doi.org/10.17743/jaes.2016.0034

optimal listening spot with that of a poor loudspeaker in an excellent reproduction room/optimal listening spot using any musical fragment. The approach thus also takes into account the well-known effect that a loudspeaker may have excellent sound quality for one type of music signal on a specific room-listening spot, while for another type of music signal on the same room-listening spot it may show a low sound quality. In order for this approach to be applicable in a wide variety of circumstances a wide quality range of different loudspeakers are used from high quality studio monitors (e.g. Quad ESL-989) to small plastic housing PC loudspeakers. Audible differences as assessed in this paper are thus extremely large and the focus of the paper is to see if a perceptual modeling approach can be used to predict these large quality differences. This perceptual approach introduces two major challenges:

1. *Ideal Reference Signal:* In the assessment of an electric input and output device, one can ask subjects to compare the reference electric input (the ideal) to the degraded electric output over a headphone. When representing the headphone in the perceptual model as a simple system with a pre-defined frequency characteristic, the model can exactly mimic the subjective test. However, it is very difficult to provide an acoustic reference signal to the subject in loudspeaker reproduction assessment that can be directly compared to the acoustic degraded loudspeaker output.
   There are two different exact reference approaches possible [26]:
   a. "here and now", where we have the illusion that the reproduced sound is present in the listening room; and
   b. "there and then", where we have the illusion that we are present in the room where the recording was made.
   Both approaches are valid hi-fi goals, but, require different recording and playback techniques. "Here and now" requires anechoic recordings that are evaluated in the listening room by playing them over the loudspeaker under test. Thus, we can directly compare the "live" signal (that was first recorded in the anechoic room) with the playback of the anechoic recording. "There and then" requires standard recordings that are reproduced over the loudspeaker to be tested and recorded in the playback room with a head and torso simulator (HATS). These HATS recordings are then compared to an ideal HATS recording of the "live" event. These recordings have to be assessed with a correct, individually-equalized headphone.
   In this research, we take a pragmatic approach, binaural recordings of the reproduced signals are made using a HATS with realistic silicon-rubber pinna and two DPA 4060 in ear microphones, while subjects judge the loudspeaker output on the same listening spot as the recordings. Reference signal recordings are made using the best quality loudspeakers available, in the ideal listening spot and in the best quality environments available. The overall sound quality of all reproduced signals are judged by subjects using the sMOS, and the recordings with the highest sMOS are taken as the reference recordings in the development of PREQUEL (see Figure 1). Note that in this approach, the subjects have no reference available and use an unknown, internal, ideal to judge the loudspeaker reproduction quality.
2. *Background Noise:* When assessing the loudspeaker reproduction in a wide variety of environments, levels of background noise will differ. While this audible background noise is only marginally taken into account by subjects in their assessment of the acoustic quality due to the effect of auditory scene analysis [27], most models are not robust against the impact of this background noise. This will be solved by introducing a noise suppression algorithm that reduces the noise found in the recorded acoustic signals.
   In order to successfully quantify the loudspeaker reproduction quality, we present a unique model baptized the Perceptual Reproduction Quality Evaluation for Loudspeakers (PREQUEL). It is based on the core elements found in the perceptual evaluation models as developed within ITU for speech [1, 2, 3, 4, 5] and music [6, 7, 8, 9] and is extended with an improved masking algorithm, based on the idea of lateral inhibition [28]. It successfully implements the solutions to the above mentioned problems and is developed on the basis of the following criteria:

   - *Overall Sound Quality:* Instead of focusing on the technical characterization of the loudspeaker system itself or on specific aspects of the acoustic output, the model quantifies the listener's quality of experience.
   - *Robustness:* The model can be used on a wide variety of loudspeakers in a wide variety of listening environments/positions.
   - *Stability:* The model accurately quantifies the sMOS of loudspeaker systems that have not been used in the training of the model.

The remainder of the paper is structured as follows. Section 1 introduces a general overview of the perceptual model, as well as the optimization of the model variables. Section 2 presents an overview of the subjective tests used to develop and validate PREQUEL. Results are presented in Section 3. Section 4 presents a discussion based on this research.

1 THE PERCEPTUAL REPRODUCTION QUALITY EVALUATION FOR LOUDSPEAKERS (PREQUEL)

A general overview of PREQUEL model can be found in Figure 2. The modeling approach largely follows the one taken in the ITU-T standardization of perceptual measurement methods for the assessment of speech quality [2, 3, 4, 5]. Each consecutive step performed by the psychoacoustic model is explained in Subsection 1.1 and each consecutive step found in the cognitive model is explained in Subsection 1.2. The PREQUEL model contains a set of variables θ, which are described in the next sec-

tions. These variables are optimized with a procedure given in Subsection 1.3. Their optimized values are provided in Table 3 and are omitted from the text. Moreover, for Equations (1)–(7) the transforms will only be provided for the reference signal $X(t)$ and its derivatives, as the transforms on degraded signal $Y(t)$ and its derivatives are identical.

1.1 Psychoacoustic Representation

Input of the Model: All signals used in this paper are in stereo and sampled at 48 kHz. Each signal has at least 1 second of silence recorded before the music fragment starts. Binaural recordings of the reference signals were made with a HATS, using the best quality loudspeakers available, in the ideal listening spot, and in the best quality environments available. The overall sound quality of the reproduced signals are judged by subjects using an sMOS scale. The reference recording with the highest sMOS is used as the input $X(t)$ for the model. The left and right channel are processed independently and their results are combined in a final degradations comparison step. The degraded signal $Y(t)$ is the binaural recording of the acoustic output of the system under test.

Calibration: The first step in the psychoacoustic model is to calibrate the level in relation to the absolute threshold (i.e. a function of frequency) by generating a sine wave with a frequency of 1000 Hz and an amplitude of 40 dB(SPL). This sine wave is transformed to the frequency domain, using a windowed Fast Fourier Transform (FFT) with a 21.34 ms frame length (1024 samples at 48 kHz sampling). The frequency axis is converted to a modified Bark scale and the peak amplitude of the resulting pitch power density is normalized to a power value of 10^4. This procedure is equivalent to the procedure used in the POLQA standard [4] and matches the sound pressure as used in the subjective experiments to the sound pressure used in the psychoacoustic model to calculate the internal representation.

The same 40 dB(SPL) sine wave is used to calibrate the psychoacoustic (Sone) loudness scale using Zwicker's law [29]. Further, the integral of the loudness density, over the Bark frequency scale, is normalized to 1 Sone.

Level Alignment: The overall power level of the reference signal $X(t)$ is scaled to match the overall power level of the degraded signal $Y(t)$.

Start Stop Indication: Recordings of the silent periods at the beginning and end of $X(t)$ and $Y(t)$ only contain background noise from the recording environment. Thus, they should be excluded in the calculations of the objective quality measurement. The model assumes a Gaussian distribution of the background noise and uses the mean $\overline{X}$ and standard deviation σ of the absolute power of the first 0.5 seconds at the start of the file as a footprint. The parts that only contain background noise are detected by sliding a frame with a size of 21.34 ms, without overlap, over $X(t)$ and $Y(t)$. The samples within this frame are considered noise if their average absolute power is within a range of 0 to 3σ of $\overline{X}$.

Further, all consecutive samples at the beginning and end of the signal that are classified as noise are cut from the signal. Subsequently the reference and degraded signals only contain the music fragment without the silence at the beginning and end of the sample. This procedure mimics the behavior of the subjects, who ignore low background noise levels in a room when they judge the loudspeaker reproduction quality [27].

Temporal Alignment: Loudspeakers do not produce time warping in their output. Thus, a simple time alignment is used that searches for a single global estimate of the delay between the reference and degraded signal. The lag is found using the cross correlation between $X(t)$ and $Y(t)$ and the aligned overlapping intervals of $X(t)$ and $Y(t)$ are used in the remainder of the pipeline.

Time Frequency Analysis: The human ear performs a time-frequency analysis. Therefore, the algorithm applies a windowed FFT with a Hamming window of 1024 samples on $X(t)$ and $Y(t)$:

$$W(T) = 0.54 - 0.46\cos(\frac{2\pi T}{1023}), T = 0 \ldots 1023, \qquad (1)$$

where $W(T)$ is the amplitude per sample. The overlap between subsequent frames is 75%. The windowed FFT results in functions of time and frequency, which are transformed into power spectra. Phase information within a single frame is discarded. The results are the power density representations $PX_{f,n}$ and $PY_{f,n}$, the power per frequency band f and frame index n. These representations are calculated for both the left and right channel of the binaural recording.

Noise Reduction: Acoustic recorded signals typically have a lot of background noise, which subjects do only marginally take into account in their assessment of the loudspeaker reproduction quality [27]. Therefore, we have to suppress this background noise. The first 0.5 seconds of the reference and degraded signals after the level alignment are classified as noise footprints. These footprints are transformed to FFT power domain. The average power of each frequency band in the reference and degraded noise footprints is calculated and subtracted from $PX_{f,n}$ and $PY_{f,n}$ respectively.

Frequency Warping: The Bark scale (i.e., the psychoacoustic equivalent of the frequency scale) models that the human hearing system has a finer frequency resolution at low frequencies, than at high frequencies [29]. This is implemented by binning consecutive frequency bands of $PX_{f,n}$ and $PY_{f,n}$, and summing their corresponding powers. The warping function that maps the frequency scale in Hertz to the pitch scale in Bark (see Table 2) approximates the values given in the literature [30]. The resulting signals $PPX_{f,n}$ and $PPY_{f,n}$ are the pitch power densities of the reference and degraded signals.

Frequency & Temporal Smearing: Masking at a perceptual level is the result of two distinct processes, a bio-mechanical and a neural process. Bio-mechanical masking is implemented using time-frequency smearing. It partially models psychoacoustic masking along the time and frequency axis and quantifies how the power from one time-

 http://dx.doi.org/10.17743/jaes.2016.0034

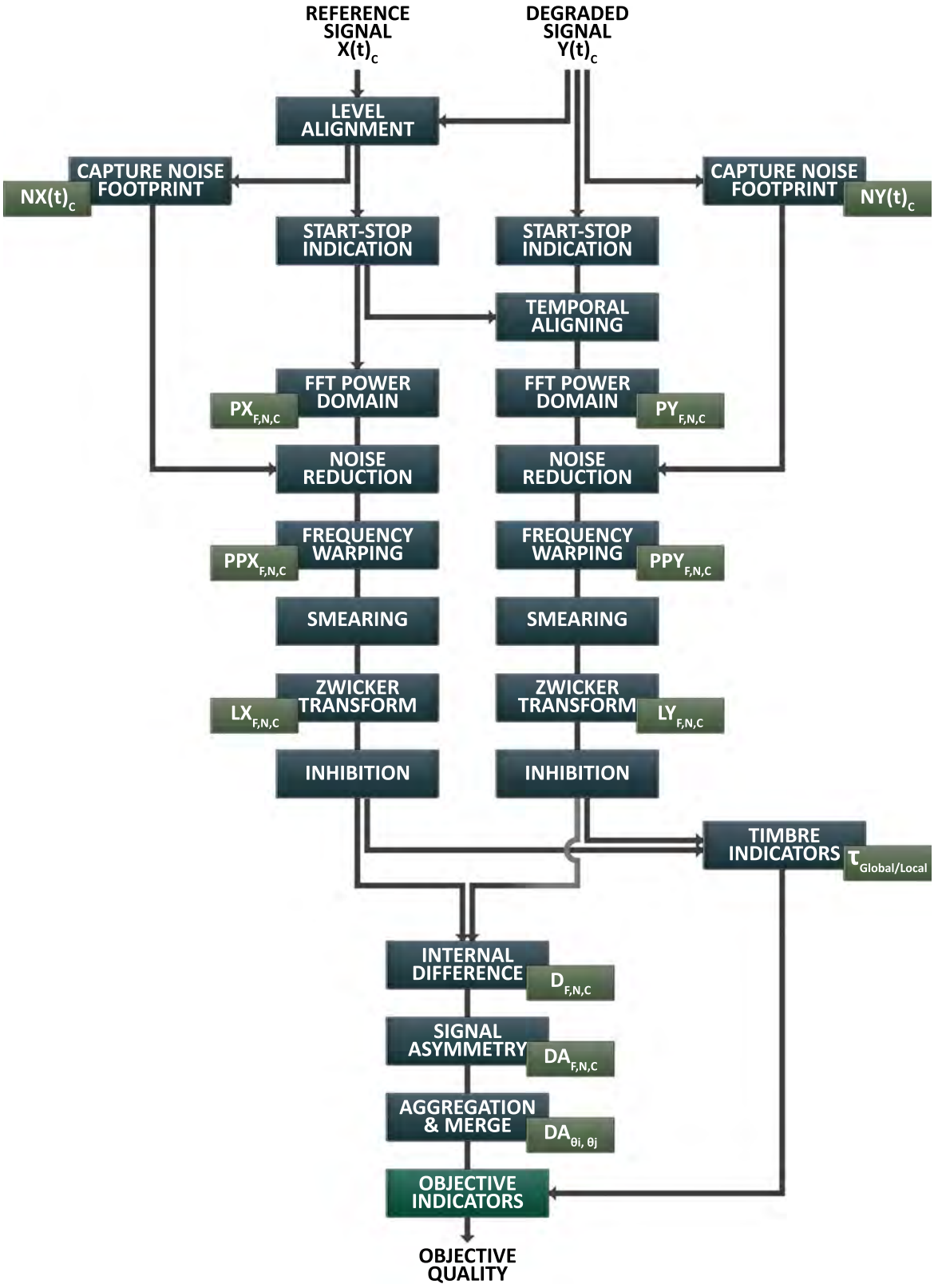

Fig. 2. The processing pipeline of PREQUEL.

frequency cell smears towards neighboring time-frequency cells [6]. Smearing in the frequency domain is applied on $PPX_{f,n}$ and $PPY_{f,n}$ as follows:

$$PPX_{f,n} = \theta_1 PPX_{f-1,n} + PPX_{f,n}. \quad (2)$$

Smearing in the time domain is applied on $PPX_{f,n}$ and $PPY_{f,n}$ as follows:

$$PPX_{f,n} = \phi(f) PPX_{f,n-1} + PPX_{f,n}, \quad (3)$$

Table 2. The warping function that maps the frequency scale in Hertz to the pitch scale in Bark (e.g. the power in every 4 consecutive bands in the range of 2,001 – 4,000 Hz are binned together).

Frequency Range (Hz)	0 – 1,000	1,001 – 2,000	2,001 – 4,000	4,001 – 8,000	8,001 - 16,000	16,001 - 24,000
Number of Consecutive Bands	1	2	4	8	16	32

where

$$\phi(f) = \begin{cases} \theta_2 & \text{if } f \geq 500; \\ -\frac{\theta_3-\theta_2}{500} f + \theta_3 & \text{otherwise.} \end{cases} \quad (4)$$

Zwicker Transformation: The reference and degraded pitch power densities are transformed to loudness densities in Sone per Bark using Zwicker's law [29]. This function transforms $PPX_{f,n}$ and $PPY_{f,n}$ to their corresponding loudness densities $LX_{f,n}$ and $LY_{f,n}$ as functions of time and frequency, as follows:

$$LX_{f,n} = S_l \left(\frac{10^{THR_f}}{0.5} \right)^{\gamma} \left[0.5 + 0.5 \frac{PPX_{f,n}}{10^{THR_f}}^{\gamma} - 1 \right], \quad (5)$$

with THR_f being the absolute threshold for the minimum audible field and γ being the Zwicker power, also denoted as optimized variable θ_4.

Frequency & Temporal Inhibition: Masking caused by inhibition at a neural level, where firing neurons suppress the firing rate of nearby neurons [28], is implemented in the loudness domain by reducing the loudness of a single time-frequency cell as a result of nearby loud time-frequency cells. Inhibition in the frequency domain is applied as follows:

$$LX_{f,n} = LX_{f,n} - \theta_5 \left(LX_{f-1,n} + LX_{f+1,n} \right) \quad (6)$$

and equivalently in time domain as:

$$LX_{f,n} = LX_{f,n} - \theta_6 LX_{f,n-1}. \quad (7)$$

Timbre Indicators: A music fragment that is reproduced with an unnatural balance between high and low frequencies leads to lower perceived sound quality [31]. This is characterized as its tone color or timbre. A first global timbre indicator is calculated using the average loudness between the low (24 Hz - θ_7 Hz) and high (θ_8 Hz - 24,000 hz) frequencies in $LX_{f,n}$ and $LY_{f,n}$, resulting in timbre values $T1X_c$ and $T1Y_c$. Subsequently, τ_1 was derived from both the left and right channel:

$$\tau_1 = \max \left(\frac{T1X_{left}}{T1Y_{left}}, \frac{T1X_{right}}{T1Y_{right}} \right). \quad (8)$$

Next, a second global timbre indicator is calculated using the average loudness between the low (24 Hz - θ_9 Hz) and high (θ_{10} Hz - 24,000 Hz) frequencies in $LX_{f,n}$ and $LY_{f,n}$, resulting in timbre values $T2X_c$ and $T2Y_c$. Subsequently, τ_2 was derived from both the left and right channel:

$$\tau_2 = \max \left(\frac{T2Y_{left}}{T2X_{left}}, \frac{T2Y_{right}}{T2X_{right}} \right) \quad (9)$$

Both τ_1 and τ_2 will be used in Section for the prediction of the sMOS.

Calculation of the Internal Difference: Two signals that only differ in overall loudness need a minimum difference in order to be discriminated. This is modeled by disturbance density $D_{f,n}$, as a function of time and frequency:

$$D_{f,n} = \begin{cases} \max(0, RD_{f,n} - M_{f,n}) & \text{if } LY_{f,n} > LX_{f,n}; \\ \max(0, RD_{f,n} - M_{f,n})\theta_{11} & \text{otherwise,} \end{cases} \quad (10)$$

with

$$M_{f,n} = \max \left(LX_{f,n}, LY_{f,n} \right) \theta_{12} \quad (11)$$

being a self-masking algorithm and

$$RD_{f,n} = |LX_{f,n} - LY_{f,n}|, \quad (12)$$

being the raw disturbance density. This algorithm pulls the raw disturbance density towards zero, which represents a dead zone (i.e., before a time-frequency cell is perceived as distorted). As such, it models the process of small time-frequency level differences being inaudible.

1.2 Cognitive Model

Asymmetry: When the system under test introduces a distortion to its input, it results in two different percepts of the output: the input signal and the distortion. However, when a distortion is introduced by leaving out a time-frequency component in the signal, this decomposition is not possible. This results in a distortion that is less objectionable, which is modeled by introducing an asymmetrical disturbance density $DA_{f,n}$:

$$D_{f,n} \left(\frac{LY_{f,n}}{LX_{f,n}} \right)^{\theta_{13}}. \quad (13)$$

Aggregation over Time and Frequency: The asymmetrical disturbance density $DA_{f,n}$ (see Equation (13)) is integrated along the frequency axis. The result is $DA_{L_i,n}$, where L_i is the L_p norm used for the frequency integration, ranging from L_1 to L_{10}. The calculation of a range of L_p norms allows to find the L_p norm that best models the effect of a local loud distortion. Such a distortion has the same average effect as a global soft distortion, which has a more severe effect than expected when applying straight forward L_1 averaging. Further, the left and right channel of $DA_{L_i,n}$ are merged by calculating the maximum disturbance over left and right in each frame n. The merged disturbance density, $MDA_{L_i,n}$, is integrated along the time axis. This results in MDA_{L_i,L_j}, where L_j is the L_p norm used for the time integration, ranging from L_1 to L_{10}. The output of the model is a vector Ω that consists of MDA_{L_i,L_j}, τ_1 and τ_2. For each music signal fragment recording k, made of a certain loudspeaker at a certain location, one can calculate an Ω_k. The set of Ω_ks and, sMOS$_k$s is used in Section 3 to predict the

 http://dx.doi.org/10.17743/jaes.2016.0034

Table 3. The 13 optimized model parameters θ. Using these values in the perceptual modeling, the maximal correlation between the predicted and the subjectively perceived sound quality is obtained.

parameter	effect	Equation	value
θ_1	frequency smearing	(2)	0.050
θ_2	time smearing	(3)	0.990
θ_3	time smearing	(4)	1.000
θ_4	Zwicker power	(5)	0.145
θ_5	time inhibition	(6)	0.300
θ_6	frequency inhibition	(7)	0.400
θ_7	timbre 1	(8)	3,400 Hz
θ_8	timbre 1	(8)	3,000 Hz
θ_9	timbre 2	(9)	1,000 Hz
θ_{10}	timbre 2	(9)	1,000 Hz
θ_{11}	self masking	(10)	0.300
θ_{12}	self masking	(11)	0.600
θ_{13}	self masking	(13)	0.100

overall sound quality of loudspeakers using multiple linear regression.

1.3 Training of the Model

The algorithm that optimizes all model variables is implemented in *C#*, and runs on a 2.4 GHz Intel(R) Core(TM) i7-3630QM CPU with 16 GB of RAM using 64-bit Windows 8.1. The optimization includes the 13 variables described in previous section and 16 additional variables that are needed to prevent instabilities of the model. Each variable is given a lower and an upper bound, based on ITU's existing perceptual evaluation models for speech [1, 2, 3, 4, 5] and music [6, 7, 8, 9]), and a Δ value that describes a finite increment of the variable. Model optimization is carried out on the basis of the correlation between the subjective MOS (sMOS) and the objective MOS predicted by the model (oMOS).

The time complexity when calculating all possible combinations of values for all variables is $\mathcal{O}(M^N)$, where N is the number of variables and M is the number of different values of each variable, based on Δ, the lower bound and the upper bound. Thus, the time to calculate the global optimum has a growth factor defined by the granularity M of the system. In order to maintain a high value of M and a time complexity independent on the exponential relation between M and N, a heuristic optimization algorithm, called Random Restart Hill Climbing [32], was implemented. While this algorithm does not guarantee that the optimal solution will be found, it is a lightweight optimization strategy that provides excellent results. The algorithm starts with a random state of variables, with values between their lower and upper bound, and iteratively attempts to find a better solution in terms of correlation between the sMOS and oMOS, by incrementally changing a single variable of the solution with its corresponding Δ. The change is accepted if the correlation coefficient of the current solution, calculated using the monotonic linear regression of the output X of the model and the sMOS of all data used in the training, is higher than the previous iteration. The search is terminated and restarted with a new random state if it stagnates over 30 iterations. So, instead of indefinitely trying to optimize a solution from one initial condition, a wider area of the solution space is searched. The search is terminated if the correlation coefficient of the best solution is above a user defined threshold. Thus, the time complexity of the algorithm is no longer dependent on the granularity of the system. Instead, it is defined as $\mathcal{O}(d)$, where d is the longest path to a solution above the given threshold.

2 SUBJECTIVE EXPERIMENTS

2.1 Experimental Setup

Although audio quality testing is best performed by running blind subjective tests [33] this is not possible when subjects have to assess audio quality in different listening rooms and listening positions. Even if a subject is moved blindfolded from one listening room to another listening room, necessary in our subjective tests, they still would detect in which room they would be seated by just listening to their own footsteps. Furthermore running tests in random order was not feasible due to limited resources. This experimental approach may result in biased judgements and thus further testing will be required to validate the model.

All tests were carried out in a stereo setup, except one surround setup where extra surround energy was created using a Yamaha V-2400-RDS surround processor. The distance between loudspeakers and subjects was adapted to the specific characteristics of the loudspeakers, large systems were assessed between 2 and 4 meters, small PC loudspeakers as close as 0.5 meters. An overview of the five different listening rooms with the loudspeaker layout and the T_{60} reverberation times is given in Figure 3.

Three experiments were run for the training and validation of PREQUEL. A total of 12 musical fragments were used that included classical large orchestras, opera/choir, solo instruments and pop/rock recordings (see Tables 4 and 5). These fragments were chosen on the basis of their high sound quality as judged by three expert listeners, all of which had more than 30 years of experience in judging acoustic recordings over loudspeaker setups. All fragments had a duration of about 30 seconds and were played consecutively with silences of four seconds between each fragment. Each fragment was individually level aligned for the optimal playback level relative to the other fragments. For acoustically recorded signals, they represent the level that was estimated to be equal to the level as experienced during the performance. For non-acoustically recorded signals, they represent the average preferred level by the subjects, ranging from rock at a level of about 90 dB(A) to solo harpsichord at a level of about 65 dB(A), fast averaging.

Table 4. The 6 music fragments used in the first experiment, a subset of the fragments as used during the development of the MPEG standard [34]. They were chosen on the basis of their high sound quality and variety in genre as judged by expert listeners in the MPEG standardization.

artist / composer	fragment
Georges Bizet	Carmen
Unknown	Trumpet solo
Tracy Chapman	Fast car
Unknown	Accordion
Unknown	Bass guitar
Unknown	Percussion

Table 5. The 6 music fragments used in the second and third experiment. They were chosen on the basis of their high sound quality and variety in genre as judged by three expert listeners.

artist / composer	fragment
Daniel Cross	The Spinroom
Stanislav Moryto	Per Uno Solo
Antonín Dvořák	Slavonic Dance Op. 72
William Walton	Set Me As a Seal Upon Thine Heart
Johann Sebastian Bach	Sarabande Partita II, BWV 826
Georg Philipp Telemann	Ach, Herr, Straf Mich Nicht In Deinem Zorn

The level differences between the loudspeakers were small (<3 dB(A)) except from musical fragments that contained a large amount of low frequency content in which cases small loudspeakers sounded significantly softer (to about 10 dB(A)). The range of dBA levels measured suggest that there were loudness differences between some of the loudspeakers, which were not equalized for the listening tests or objective measurements. These loudness differences largely arise out of frequency response differences between devices. However, equalizing overall loudness could have given rise to unrealistic mid-range level differences. Thus, on the one hand, loudness differences leaked into the quality judgement predominantly as a side effect of the big differences between the low frequency response of the loudspeakers. On the other hand, we have assured an ecologically valid approach, which was one of our key aims.

All fragments in each experiment were binaurally recorded for processing by the PREQUEL model, using a HATS. Each individual experiment used six subjects who judged a sequence of six music fragments.

Each experiment was performed by six subjects using naive and expert listeners. A total of 18 subjects were used, consisting of 16 males and 2 females, with an age ranging from 22 to 74. The subjects were not screened for possible hearing loss. Subjects were instructed to judge the overall sound quality produced by several loudspeaker reproduction systems relative to each other. Note that the subjects had no direct "ideal" reference available and used an unknown, internal, ideal to judge the loudspeaker reproduction quality. An evaluation scale based on the Netherland's school reporting system was used for the judgments ranging from 1 ("bad") to 10 ("excellent") (cf. [35]). As such, this scale was the best choice for the participants to express their quality opinion. Before the experiment started subjects had a training session that provided them a direct comparison of music fragments played over any system they would like to hear. During the test any direct comparison was made available upon request by a subject. This direct comparison, combined with the fixed order of play out; allowed them to develop a stable "ideal" reference and stable quality judgements on all systems. Five different rooms were used in the three subjective tests, see Figure 3 for an overview of the properties of these rooms. One subjective test typically lasted about six hours.

The first two experiments were performed using two high quality with excellent acoustic properties (see Figures 3A and 3B). Subjects were seated at 2, 3 or 4 different positions in each room and judged the quality of 6 loudspeaker systems, ranging from high quality studio monitors, including both electrodynamic and electrostatic loudspeakers, a high quality omnidirectional loudspeaker system to very low quality PC loudspeakers (see Tables 6 and 7). All loudspeaker systems were assessed using the optimal on axis position as well as up to 3 different off axis positions. All recordings used in the objective assessment were made at exactly the same spots where the subjects judged the audio quality. There were a total of 36 different loudspeaker reproduction evaluation setups, which resulted in a total of 216 fragments that had to be judged.

The third experiment was performed using three average to low quality listening rooms (see Figure 3C-E). Subjects judged 11 loudspeaker systems in the 3 listening rooms at 2 different positions. The quality of the systems ranged from a high quality omnidirectional system (electrodynamic), to very low quality consumer type loudspeakers. Furthermore, the experiment included a 4-channel surround system and a number of 2-channel room reverberation algorithms (see Table 8).

There were a total of 22 different loudspeaker reproduction evaluation setups, which resulted in a total of 132 fragments that had to be judged. Note that in each experiment listeners were rating the combined quality effect of the loudspeaker, the room, the spatial reproduction mode, and any spatial processing employed.

2.2 Subjective Results

The subjective results consisted of three databases with entries that represent a musical fragment recording of a certain loudspeaker/positioning and its associated opinions of the subjects in terms of sMOS. This section discusses the analysis of these *raw* sMOS values for each of the three databases. To ensure the ecological validity of these analysis, we applied a baseline-free analysis, lacking data nor-

 http://dx.doi.org/10.17743/jaes.2016.0034

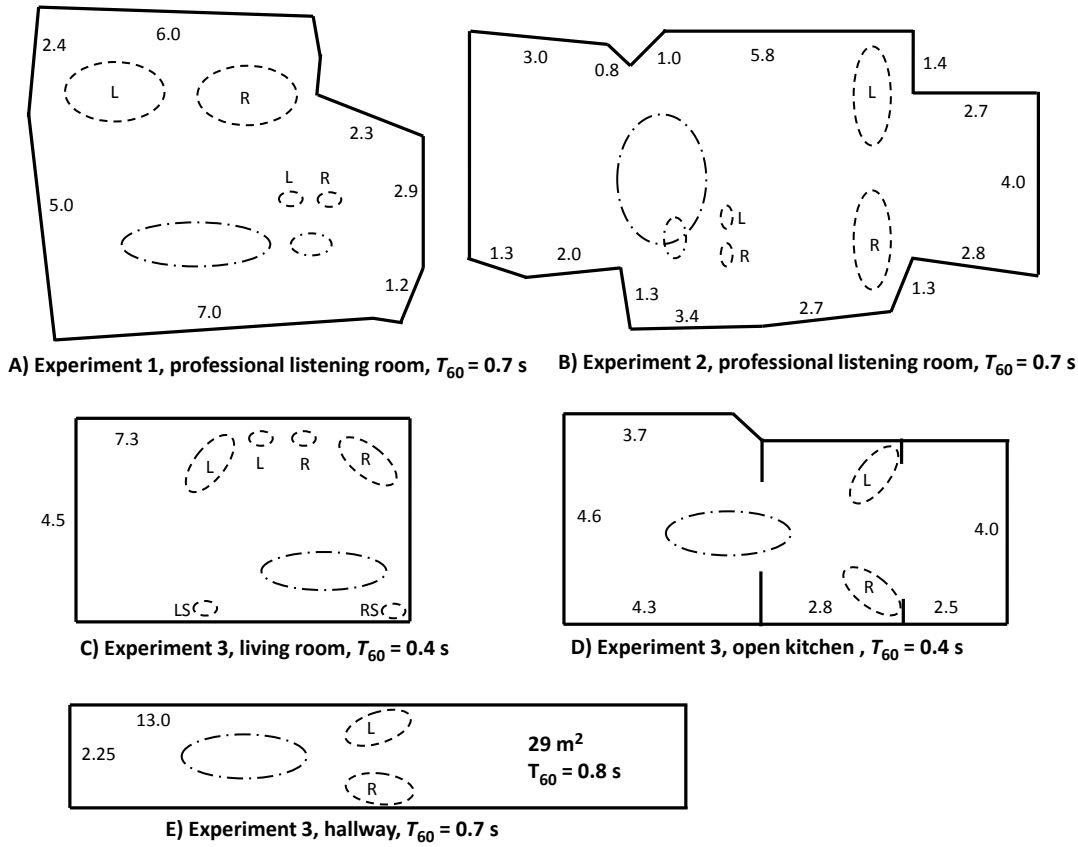

Fig. 3. Layout and reverberation time T_{60} of the five different listening rooms used in the three subjective experiments. All dimensions are in meters, the dashed areas represent Left (L), Right (R), Left Surround (LS) and Right Surround (RS) loudspeaker positions. Dash-dotted areas represent the listening positions used in the subjective tests (on- and off-axis).

malization (cf. [35, 36]), which can be conveniently applied in real-life situations [37]. First, the statistical considerations taken in the analysis will be presented. Next, in line with the main aims of this study, we will analyze the subjective MOS (sMOS) scores, using an univariate repeated measures analysis of variance (ANOVA) (Huynh-Feldt). The listeners sMOS scores served as dependent variable and music style, listening position, and the loudspeaker, the room, the spatial reproduction mode and any spatial processing employed served as independent variables. This will be done for each database separately after which general conclusions are drawn.

All three tests are reported with their F-values, degrees of freedom, power, and level of significance. If the level of significance is close to zero, this is denoted with $p < .001$, instead of providing an exact statistic. As measure of effect size partial eta squared (η^2) is reported to indicate the proportion of variance accounted for (i.e., a generalization of r/r^2 and R/R^2 in correlation/regression analysis) [38]. The threshold for reporting results is $p \leq .010$. Effect size and the amount of explained variance are interpreted in line with Cohen's suggestions [39] (p. 157). On average, interaction effects between factors explained $\pm 40\%$ of the variance the factors did separately (range $eta^2 : .075 - -.416$). Hence, the interaction effects did not provide additional information; therefore, interaction effect have been omitted. Moreover, results of pairwise tests between the music style, listening position, and loudspeakers (and listening rooms) also have been omitted, as these only provided the obvious significant effects (e.g., PC speakers received lower sMOS scores than high end speakers), due to a lack of data samples.

Database 1: Music was shown to have a significant, medium effect on the listeners' sMOS scores, $F(5,25) = 3.167, p = .024, \eta^2 = .388$. However, both the listening position ($F(1,5) = 81.029, p < .001, \eta^2 = .942$) and the loudspeakers ($F(5,25) = 45.026, p < .001, \eta^2 = .900$) were shown to be a much more powerful predictor of the sMOS scores. The listeners' sMOS scores correlated very high with each other (between $r = .649$ and $r = .964$ (Pearson), see also Table 9).

Database 2: Neither music ($F(5,25) = 2.726, p = .097, \eta^2 = .353$) nor position ($F(3,15) = 2.491, p = .130, \eta^2 = .333$) were shown to have a significant impact on the listeners' sMOS scores. However, with both aspects the η^2 indicates medium effect size that can be accounted for. The loudspeakers did have a significant impact on the listeners' sMOS scores ($F(5,25) = 58.824, p < .001, \eta^2 = .920$) and were shown to be powerful predictor of the sMOS scores. The listeners' sMOS scores correlated very high with each other (between $r = .651$ and $r = .894$ (Pearson), see also Table 10).

Table 6. The 6 loudspeaker systems used in the first experiment. Their quality ranges from a high quality studio monitor and a high quality omnidirectional system (radiating over 360° in the lateral plane) to average consumer type systems, including a very low quality PC loudspeaker.

Loudspeaker	Type of reproduction
Bloomline	2 channel, 4x electrodynamic, omnidirectional
KEF Corelli	2 channel, 2x electrodynamic, front radiating
Home build	2 channel, 2x electrodynamic, front and rear radiating
Aldi Life 5454	2 channel, 2x electrodynamic, front radiating
Grimm Audio LS1	2 channel, 2x electrodynamic, front radiating
PC Mini Speaker	2 channel, 2x active electrodynamic, front radiating

Table 9. Listeners' consensus table for database 1, including their average. In all cases, the Pearson correlation was highly significant (i.e., $p < .001$).

	1	2	3	4	5	6	ave.
1		.731	.725	.760	.834	.786	.881
2			.869	.834	.881	.793	.929
3				.905	.852	.693	.921
4					.926	.649	.933
5						.771	.964
6							.849

Table 10. Listeners' consensus table for database 2, including their average. In all cases, the Pearson correlation was highly significant (i.e., $p < .001$).

	1	2	3	4	5	6	ave.
1		.894	.766	.812	.660	.869	.913
2			.816	.826	.723	.892	.941
3				.747	.651	.867	.889
4					.753	.827	.915
5						.810	.840
6							.963

Table 11. Listeners' consensus table for database 3, including their average. In all cases, the Pearson correlation was highly significant (i.e., $p < .001$).

	1	2	3	4	5	6	ave.
1		.709	.701	.721	.705	.791	.907
2			.778	.652	.431	.617	.828
3				.719	.512	.646	.857
4					.598	.714	.866
5						.684	.770
6							.867

Database 3: Music were shown to have a significant, medium effect on the listeners' sMOS scores, $F(5,25) = 3.484, p = .027, \eta^2 = .411$. Listening position did not have a significant effect on the sMOS scores given, $F(1,5) = 1.534, p = .271, \eta^2 = .235$. The loudspeaker-room combination were shown to be a powerful predictor of the sMOS scores, $F(10,50) = 27.084, p < .001, \eta^2 = .844$. The listeners' sMOS scores correlated medium to high with each other (between $r = .431$ and $r = .791$ (Pearson), see also Table 11).

Conclusions: Music did have an influence on the sMOS scores. However, this was shown only to be significant in two of the three databases. Only for one of the three databases, position was shown to have a significant influence on the sMOS. This warrants additional research to unveil the influence of listening position across music styles, listening rooms, and speakers. With all three data sets, the speaker-room combinations did have a very strong influence on the sMOS scores. Taken together, the three databases each contain substantial variance in their sMOS scores, which can be contributed to the three factors influenced. Moreover, the three databases together also include a substantial variance in listening rooms, ranging from two recording studios to three typical consumer listening rooms (i.e., living room, kitchen, and hallway). Given the variance within the complete set of data, the listeners showed a remarkably high consensus in their sMOS scores.

Tables 9–11 show that all subjects have a Pearson correlation of at least 0.77 (one of the naive subjects) between their individual opinion and the average opinion of the group, while the best subject (one of the experts) has a correlation of about 0.96. Thus, despite the fact that subjects mentioned that they had issues regarding the difficulty of taking into account all possible degradation parameters (e.g., timbre, envelopment, localization and room resonances), the consistency in judgement is very high, verifying the high relevance of the subjective data when describing the overall perceived sound quality.

Taken together listeners agree in their judgement on music across a wide range of circumstance despite their inevitable intra and inter-individual differences [36]. This suggests that the quantification of loudspeakers' perceived sound quality by assessing their acoustic output using a set of music signals is indeed possible. Hence, the development of the proposed generic PREQUEL, a unique, robust psychoacoustic model of listener's percept of loudspeakers' sound quality, should be feasible.

Table 7. The 6 loudspeakers systems used in the second experiment. Their quality ranges from high quality studio monitors (electrodynamic and electrostatic) and a high quality omnidirectional system (radiating over 360° in the lateral plane) to average consumer type systems, including a very low quality PC loudspeaker.

Loudspeaker	Type of reproduction
Bloomline	2 channel, 4x electrodynamic, omnidirectional
KEF Corelli	2 channel, 2x electrodynamic, front radiating
Quad ESL-989	2 channel, 2x electrostatic, front and rear radiating
Tannoy Small Studio Monitor	2 channel, 2x electrodynamic, front radiating
Tannoy Large Studio Monitor	2 channel, 2x electrodynamic, front radiating
PC Mini Speaker	2 channel, 2x Active electrodynamic, front radiating

3 MODELING RESULTS

3.1 Scale Normalization

In order to obtain a wide acceptance of the modeling results the final subjective sMOS scores used in the model development are normalized per experiment in two subsequent steps. First, a Z transform is applied:

$$X' = \frac{X - \mu}{\sigma}, \tag{14}$$

with X the original sMOS score, μ the mean, and σ the standard deviation of the data of each subject, providing X', the normalized sMOS score. Next, the X'-score of each fragment is transformed to the standard ITU five point scale [10] (also see Table 1), for each experiment individually, as follows:

$$X'' = \frac{4(X' - \min\{X'\})}{\max\{X'\} - \min\{X'\}} + 1, \tag{15}$$

with $\min\{X'\}$ and $\max\{X'\}$ being respectively the lowest and the highest X' score in each experiment. Consequently, the results presented next are based on these X'', the adjusted, normalized ITU standard sMOS scores.

3.2 Training and Validation

The three experiments described in Section 2 were used to create three databases, each consisting of a collection of binaurally recorded music signals and their normalized sMOS (also see Subsection 3.1). As explained in the introduction, for each music fragment the recording with the highest sMOS in the database is used as the reference recording $X(t)$ in the modeling. Furthermore, each database is split in two equal parts: a training part and a validation part. This splitting is carried out on the basis of four equal sMOS intervals where both the training and validation set contain an equal amount of sMOS values in each sMOS interval. Within each interval the training and validation set were chosen randomly. This method prevents overtraining and enforces a balanced training and validation set where both have a full range of sMOS values. The performance of the model is measured in terms of the Pearson correlation coefficient r between the values Ω_k of each fragment, as introduced in section 1.2, and its corresponding normalized sMOS_k values, using a monotonic polynomial regression.

Training: The training set is used to develop a robust model that is trained context independently in such a way that it is able to quantify the subjective sound quality of loudspeaker systems. First, all model variables θ, τ and the L_p norms $L1_{freq}$, $L1_{time}$, $L2_{freq}$ and $L2_{time}$ are trained (see Subsection 1.3), resulting in the optimized variables that are the same for all databases. This allows the calculation of the following single value ψ for each binaurally recorded music signal:

$$\begin{aligned}\psi = &-\left[MDA_{L1_{freq},L1_{time}}\tau_1\right] \\ &-0.045\left[MDA^2_{L2_{freq},L2_{time}} + \tau_1\tau_2\right] \\ &+0.034\left[\tau_2 + \tau_2^{-1}\right].\end{aligned} \tag{16}$$

In general context effects in subjective experiments can to lead to a different quality judgement of the same degradation, e.g. due to voting preferences or the balance of conditions. This problem can be illustrated with the following example. Assume that we have a robust model that gives an average objective measurement score of 2.5 on the ITU scale to a signal. Next, this signal is presented in two different experiments. In the first experiment, this signal has the highest quality of all degraded signals while in the second experiment it has the lowest quality of all degraded signals. This results in two different sMOS given by subjects in the two experiments while the degraded signal in both experiments is exactly the same [40].

The solution to this context problem is to use a different 3^{rd} order monotonic polynomial regressions for each database. This procedure is in line with the procedure followed in the ITU standardization [1, 3, 5, 8] and results in a robust model that can predict the sMOS context independently. The correlations of the multiple polynomial regression for the trained parts of the three databases are 0.94, 0.90, and 0.88 per loudspeaker per individual listening position averaged over all music fragments. This regression uses the same optimized values for θ and the same nonlinear function to calculate ψ_k for all data in A, while using a context dependent polynomial P_m for each database, where m is the index of each database. So a different polynomial P is used for training the model to fit each of the three data

Table 8. The 11 loudspeaker systems used in three rooms in the third experiment. Their quality ranges from a high quality omnidirectional system (radiating over 360° in the lateral plane), to average and low quality consumer type systems.

#	Room	Loudspeaker	Type of reproduction
1	Living room	Canton mini modified	2 channel, 2x electrodynamic, front radiating
2		Canton mini modified + Yamaha Room Simulator Hall Vienna	2 channel, 2x electrodynamic, front radiating
3		Boston CR65	2 channel, 2x electrodynamic, front radiating
4	Kitchen	Bloomline	2 channel, 4x electrodynamic, omnidirectional
5		Tetra Home Build + Aldi + Yamaha Surround Field Filler	4 channel, 4x electrodynamic, 2x front radiating and 2x rear radiating RS and LS
6		Tetra Home Build + Aldi + Yamaha Room Simulator Hall Vienna	4 channel, 4x electrodynamic, 2x front radiating and 2x rear radiating RS and LS
7		Samsung TV	2 channel, 2x electrodynamic, front radiating
8	Hallway	KEF Corelli	2 channel, 2x electrodynamic, front radiating
9		Bowers & Wilkens DM6	2 channel, 2x electrodynamic, front radiating
10		Bowers & Wilkens DM6 + Yamaha Room Simulator Hall Vienna	2 channel, 2x electrodynamic, front radiating
11		Cheap Brand	2 channel, 2x electrodynamic, front radiating

sets, while in the validation the polynomial associated with the database from which the validation set was taken, is used.

The model should be able to predict the quality of a wide variety of loudspeakers in a wide variety of environments. Thus, the optimized variables θ, the nonlinear function used to calculate ψ_k, and the monotonic polynomial P must be identical for all data processed by the model. These constraints guarantee that the context of each experiment does not influence the performance of the model.

Validation: The validation of the model is performed using a blind prediction of the sMOS on the signals in the validation set. First a ψ_k is calculated with Equation (16), using the trained θ, τ and the L_p norms, for each recorded signal in the validation set. Next the three polynomials associated with the training part of the three databases are applied to the unseen parts of the data sets to validate PREQUEL. The results can be found in Figures 4, 5, and 6 for each database separately. The figures show the 95% confidence intervals for the normalized sMOS values and the ideal linear regression $Y = X$. The resulting correlations between PREQUEL's predictions and the normalized sMOS per loudspeaker for the unseen validation parts of the databases are 0.86, 0.92, and 0.78, per individual listening position, averaged over all music fragments.

4 DISCUSSION

This paper presents a unique loudspeaker reproduction quality measurement model baptized Perceptual Reproduction Quality Evaluation for Loudspeakers (PREQUEL). Whereas previous research focused on the quantification of the loudspeaker system itself or on specific technical aspects of the acoustic output of loudspeakers, this paper focuses on listeners' overall perceived sound quality of individual loudspeakers in a wide variety of listening environments, using a large and diverse set of music fragments. The major difference with classical technical measurement approaches is that one does not need any technical measurements on the loudspeaker and or listening

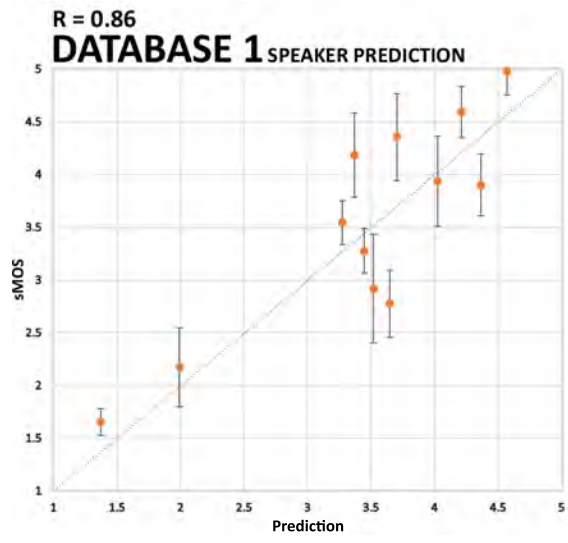

Fig. 4. The blind prediction for part B of database DB_1, plotted per loudspeaker per individual listening position averaged over all music fragments. The graph shows the 95% confidence interval of the subjective data and the ideal linear regression $Y = X$. The correlation coefficient between the predicted values and the sMOS is 0.86.

 http://dx.doi.org/10.17743/jaes.2016.0034

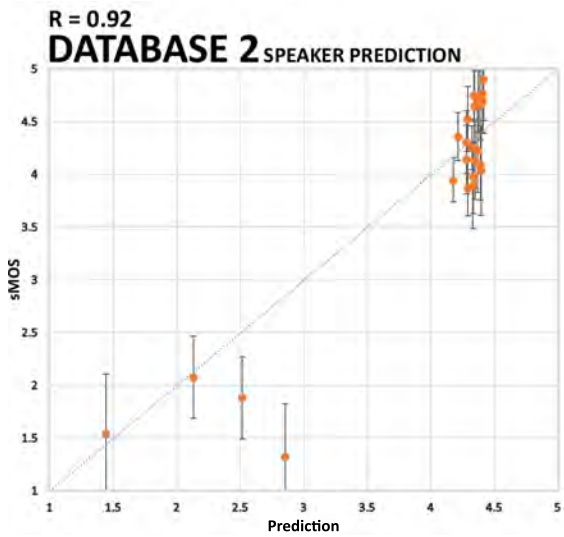

Fig. 5. The blind prediction for part B of database DB_2, plotted per loudspeaker per individual listening position averaged over all music fragments. The graph shows the 95% confidence interval of the subjective data and the ideal linear regression $Y = X$. The correlation coefficient between the predictor values and the sMOS is 0.92. The severe outlier is a small PC loudspeaker judged at close distance.

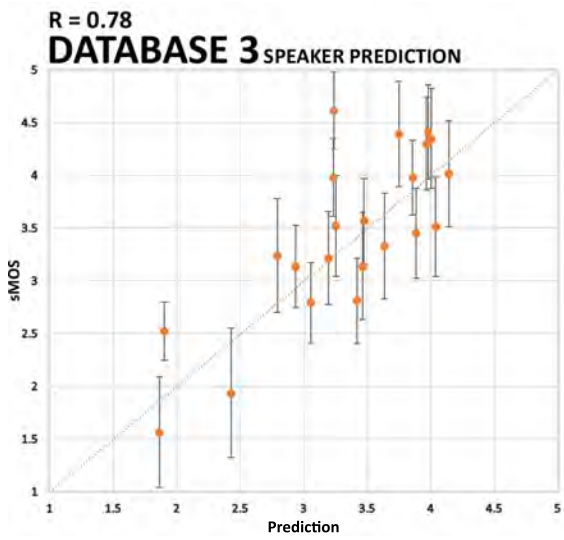

Fig. 6. The blind prediction for part B of database DB_3, plotted per loudspeaker per individual listening position averaged over all music fragments. The graph shows the 95% confidence interval of the subjective data and the ideal linear regression $Y = X$. The correlation coefficient between the predictor values and the sMOS is 0.78.

room, only recordings of musical fragments played over the loudspeakers in the room for which one wants to assess the quality, are used. This new perceptual modeling approach thus allows a direct comparison between the perceived quality of an excellent loudspeaker in a bad reproduction room/non-optimal listening spot with that of a poor loudspeaker in an excellent reproduction room/optimal listening spot using any musical fragment. The approach thus also takes into account the well-known effect that a loudspeaker may have excellent sound quality for one type of music signal on a specific room-listening spot, while for another type of music signal on the same room-listening spot it may show a low sound quality.

PREQUEL's perceptual measurement approach introduces two major challenges:

1. Subjects cannot be provided with an acoustic reference signal and, consequently, base their judgement on an unknown reference. However, a perceptual model needs a known reference signal in order to be able to predict the perceived quality. We created these reference signals by making binaural recordings with a Head and Torso Simulator (HATS), using the best quality loudspeakers available, in the ideal listening spot in the best quality listening environments available. Then, the reference signal with the highest subjective quality given by subjects (i.e., the sMOS) is compared to the acoustic degraded loudspeaker output.
2. Handling different levels of background noise in the listening environments. This is solved by introducing a noise suppression algorithm, which operates on both the reference and degraded signals.

The solutions to these challenges led to the development of the PREQUEL model, which uses a perceptual model to predict the subjective quality ratings of loudspeaker reproduction systems. Consistency checks performed on the subjective data show correlations of 0.77 and higher between listeners' individual opinion and the average opinion of the group. Hence, the listeners very much agreed on the overall perceived sound quality. Nevertheless, simultaneously, it shows that true ground truth is perhaps not possible, with such a significant variance within the group of listeners (i.e., so-called interpersonal variance, where intra personal variance can be a challenge as well) [41, 36].

Three databases were created for the training and validation of PREQUEL. Using half of this available data, PREQUEL is trained in such a way that it is able to accurately predict the subjective quality ratings. The other unseen half of the data is used to validate PREQUEL, which showed that it is promising model to quantify the sound quality of a wide range of loudspeakers systems in a wide variety of listening rooms and listening positions. A high correlation was achieved for both the training ($r = 0.91$) and validation ($r = 0.86$) phases. Please note that for two of the three databases these results were obtained without measurements roughly in the middle of the scale.

Follow-up studies should address further model validations. Possibly such studies could also come up with solutions for the difficulty to carry out blind subjective tests in the context of the perceptual modeling approach. Furthermore, new subjective tests could use the combined effect of loudspeaker distortions and other types of distortions such as amplitude clipping in amplifiers, low bit rate audio coding and time clipping in packet switching networks. The combined effect assessment allows to run double blind tests over a limited set of degradations. It is expected that PREQUEL's performance will drop when new subjective data becomes available. However, PREQUEL allows adap-

tation (or retraining), allowing it to cope with these new loudspeakers, rooms, and spatial reproduction modes as well as types of distortion. Additionally, this process may unveil these aspects' relative importance.

Distinct implementations of the HATS will also impact PREQUEL's performance, as it is currently tailored to one specific HATS. Possible influential parameters are the size of the ears, the size of the head, and the difference in quality of the microphones used in the HATS. Especially the size of the ears and head have a significant impact on the head related transfer functions (HRTFs) and, thus, possibly on the model development. However, as with other sources of variance PREQUEL can be retrained.

Taken together, this paper introduces a HATS-based model that enables to predict the reproduction quality for loudspeakers taking into account the impact of the reproduction room. It quantifies the loudspeakers' overall perceived sound quality by assessing their recorded acoustic output using a set of music signals in combination with a perceptual model. It is founded on a set of three databases, which include listener's subjective judgements of music fragments in a range of listening rooms and listening positions reproduced over a wide variety of loudspeakers. The key difference with classical measurement approaches is that is does not require any technical measurements on either the loudspeaker or listening room.

5 ACKNOWLEDGEMENTS

The authors would like to thank Leo de Klerk (Bloomline), Martin Goosen (Bloomline / DPA) and Eelco Grimm (Grimm Audio) for providing two high quality listening rooms, several loudspeaker systems and the HATS. Moreover, the authors thank the anonymous experts for their thorough and constructive reviews, which helped to improve the initial manuscript significantly.

6 REFERENCES

[1] John G Beerends and Jan A Stemerdink. A perceptual speech-quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 42(3):115–123, 1994.

[2] John G Beerends, Andries P Hekstra, Antony W Rix, and Michael P Hollier. Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model. *Journal of the Audio Engineering Society*, 50(10):765–778, 2002.

[3] ITU-T. Recommendation p.862: Perceptual Evaluation of Speech Quality (PESQ) – an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *International Telecommunication Union, Geneva*, 2001.

[4] John G Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl. Perceptual Objective Listening Quality Assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement. Part II – Perceptual model. *Journal of the Audio Engineering Society*, 61(6):366–384, 2013.

[5] ITU-T. Recommendation p.863: Perceptual Objective Listening Quality Assessment (POLQA). *International Telecommunication Union, Geneva*, 2014.

[6] John G Beerends and Jan A Stemerdink. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40(12):963–978, 1992.

[7] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes. PEAQ – The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, 2000.

[8] ITU-R. Recommendation bs.1387: Method for objective measurements of perceived audio quality. *International Telecommunication Union, Geneva*, 1996.

[9] Peter Počta and John G Beerends. Subjective and objective assessment of perceived audio quality of current digital audio broadcasting systems and web-casting applications. *IEEE Transactions on Broadcasting*, 61(3):407–415, 2015.

[10] ITU-T. Recommendation p.800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva*, 1996.

[11] Floyd E Toole. Subjective measurements of loudspeaker sound quality and listener performance. *Journal of the Audio Engineering Society*, 33(1/2):2–32, 1985.

[12] Floyd E Toole. Loudspeaker measurements and their relationship to listener preferences: Part 1. *Journal of the audio Engineering Society*, 34(4):227–235, 1986.

[13] Floyd E Toole. Loudspeaker measurements and their relationship to listener preferences: Part 2. *Journal of the Audio Engineering Society*, 34(5):323–348, 1986.

[14] Floyd E Toole. *Sound reproduction: loudspeakers and rooms*. Oxford, UK: Focal Press / Elsevier, 2008.

[15] Wolfgang Klippel and Ulf Seidel. Measurement of impulsive distortion, rub and buzz and other disturbances. In *Proceedings of the 114th Convention of the Audio Engineering Society*, page paper 5734, 2003.

[16] Yasuaki Tannaka and Tsuneji Koshikawa. Correlations between sound field characteristics and subjective ratings on reproduced music sound quality. *Journal of the Acoustical Society of America*, 86(2):603–620, 1989.

[17] Wolfgang Klippel. Assessing the subjectively perceived loudspeaker quality on the basis of objective parameters. In *Audio Engineering Society Convention*, volume 88, page paper 2929, Montreux, Switzerland, 13–16 March 1990. New York, NY, USA: Audio Engineering Society, Inc.

[18] Sean E. Olive. A multiple regression model for predicting loudspeaker preference using objective measurements: Part I – listening test results. In *Audio Engineering Society Convention*, volume 116, page paper 6113, Berlin, Germany, 8–11 May 2004. New York, NY, USA: Audio Engineering Society, Inc.

[19] Sean E. Olive. A multiple regression model for predicting loudspeaker preference using objective measurements: Part II – development of the model. In *Audio Engi-*

neering Society Convention, volume 116, page paper 6190, Berlin, Germany, 8–11 May 2004. New York, NY, USA: Audio Engineering Society, Inc.

[20] Chin-Tuan Tan, Brian CJ Moore, and Nick Zacharov. The effect of nonlinear distortion on the perceived quality of music and speech signals. *Journal of the Audio Engineering Society*, 51(11):1012–1031, 2003.

[21] Chin-Tuan Tan, Brian CJ Moore, Nick Zacharov, and Ville-Veikko Mattila. Predicting the perceived quality of nonlinearly distorted music and speech signals. *Journal of the Audio Engineering Society*, 52(7/8):699–711, 2004.

[22] Alf Gabrielsson, Ulf Rosenberg, and Håkan Sjögren. Judgments and dimension analyses of perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America*, 55(4):854–861, 1974.

[23] Alf Gabrielsson and Håkan Sjögren. Perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America*, 65(4):1019–1033, 1979.

[24] Robert Conetta, Tim Brookes, Francis Rumsey, Slawomir Zielinski, Martin Dewhirst, Philip Jackson, Søren Bech, David Meares, and Sunish George. Spatial audio quality perception (Part 1): Impact of commonly encountered processes. *Journal of the Audio Engineering Society*, 62(12):831–846, 2015.

[25] Robert Conetta, Tim Brookes, Francis Rumsey, Slawomir Zielinski, Martin Dewhirst, Philip Jackson, Søren Bech, David Meares, and Sunish George. Spatial audio quality perception (Part 2): A linear regression model. *Journal of the Audio Engineering Society*, 62(12):847–860, 2015.

[26] Stanley P Lipshitz. Stereo microphone techniques: Are the purists wrong? *Journal of the Audio Engineering Society*, 34(9):716–744, 1986.

[27] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA, USA: The MIT Press, 1994.

[28] T. Houtgast. Psychophysical evidence for lateral inhibition in hearing. *Journal of the Acoustical Society of America*, 51(6B):1885–1894, 1972.

[29] Eberhard Zwicker and Richard Feldtkeller. *Das Ohr als Nachrichtenempfänger*. Stuttgart, Germany: S. Hirzel Verlag, 2nd edition, 1967.

[30] Eberhard Zwicker. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, 33(2):248, 1961.

[31] Fares El-Azm and Jan Abildgaard Pedersen. Natural timbre in room correction systems (Part II). In *The Proceedings of the AES 32nd International Conference: DSP for loudspeakers*, pages 21–29, Hillerod, Denmark, 21–23 September 2007. New York, NY, USA: Audio Engineering Society, Inc.

[32] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall series in Artificial Intelligence. Upper Saddle River, NJ, USA: Pearson Education, Inc., 3rd edition, 2010.

[33] Floyd E. Toole and Sean Olive. Hearing is believing vs. believing is hearing: Blind vs. sighted listening tests, and other interesting things. In *Audio Engineering Society Convention*, volume 97, page paper 3894, San Franciso, CA, USA, 10–13 November 1994. New York, NY, USA: Audio Engineering Society, Inc.

[34] ISO/IEC JTC1/SC2/WG11-MPEG. MPEG/Audio test report. Technical Report MPEG90/N0030, MPEG, Stockholm, Sweden, October 1990.

[35] Egon L. van den Broek, Peter M. F. Kisters, and Louis G. Vuurpijl. Content-based image retrieval benchmarking: Utilizing color categories and color distributions. *Journal of Imaging Science and Technology*, 49(3):293–301, 2005.

[36] Egon L. van den Broek. *Affective Signal Processing (ASP): Unraveling the mystery of emotions*. PhD thesis, Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede, The Netherlands, 2011.

[37] Egon L. van den Broek and Joyce H. D. M. Westerink. Considerations for emotion-aware consumer products. *Applied Ergonomics*, 40(6):1055–1064, 2009.

[38] E. C. Young. *Vector and tensor analysis*. Pure and Applied Mathematics: A series of monographs and textbooks. New York, NY, USA: Marcel Dekker, Inc., 2nd edition, 1993.

[39] J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

[40] Slawomir Zielinski, Francis Rumsey, and Søren Bech. On some biases encountered in modern audio quality listening tests – a review. *Journal of the Audio Engineering Society*, 56(6):427–451, 2008.

[41] E. L. van den Broek. Beyond biometrics. *Procedia Computer Science*, 1(1):2505–2513, 2010.

THE AUTHORS

John G. Beerends

Kevin van Nieuwenhuizen

Egon L. van den Broek

John G. Beerends received a BEng degree in electrical engineering from the HTS (Polytechnic Institute) of The Hague, The Netherlands, in 1975, the MSc degree from the University of Leiden in 1984. In 1983 he was awarded a prize of DFl 45000 by Job Creation for the further development of his patented asymmetric loudspeaker enclosure design. From 1984 to 1989 he worked at the Institute for Perception Research where he received a PhD from the Technical University of Eindhoven in 1989. The main part of his doctoral work, which deals with pitch perception, was published in the Journal of the Acoustical Society of America. The results of this work led to a patent on a pitch meter by the N.V. Philips Gloeilampenfabriek. From 1986 to 1988 he worked on a psycho-acoustically optimized loudspeaker system for the Dutch loudspeaker manufacturer BNS. The system was introduced at the Dutch consumer exhibition FIRATO in 1988. In 1989 he joined the KPN Research where he worked on audio and video quality assessment, audio-visual interaction, and on audio coding (speech and music). This work led to several patents and two measurement methods for objective, perceptual, assessment of audio quality which he developed together with Jan Stemerdink. The first one dealt with telephone-band speech and was standardized in 1996 as ITU-T Recommendation 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). Most of the work on audio quality (speech, music and audiovisual interaction) was published within the Audio Engineering Society and the ITU. From 1996 to 2002 he worked with Andries Hekstra 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 focussed 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 January 2003 he joined TNO, which took over the research activities from KPN, where he worked on the objective measurement of speech intelligibility, (super) wideband speech quality, degradation decomposition, hearing aid quality, videophone quality and data chirping techniques. The main focus was on speech quality and intelligibility assessment for the normal hearing and hearing impaired. In the period 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 (Perceptual Objective Listening Quality Assessment, POLQA) in 2011. Currently he is working on extending the perceptual measurement approach towards acoustic domain measurements (loudspeaker reproduction quality, including the impact of the reproduction room) and on the glass box modeling of audio, speech, video, data services. John Beerends is author of more than 100 (conference) papers/ITU contributions and 35 patents. In 2003, he received an AES fellowship award for his work on audio and video quality measurement.

●

Kevin van Nieuwenhuizen received a BSc degree and MSc degree, both cum laude, in computer science from the Utrecht University, The Netherlands, in 2013 and 2015 respectively. He specialized in the field of gaming and media technology and has won several awards at different game development competitions. Kevin worked on several small gaming projects and currently develops, and maintains, pension related applications using the latest Microsoft technologies.

●

Egon L. van den Broek received a MSc in Artificial Intelligence (AI) (2001), his first PhD in Social Sciences (2005), and his second PhD in Electrical Engineering, Mathematics, and Computer Science (2011). Currently, he is assistant professor and research director of the Center for Research on data-driven User eXperience (CRUX) at the Utrecht University, founding partner at Information eXperience (IX) BV, and consultant (e.g., for TNO, Philips, and the United Nations). His interests are on pattern recognition, interaction technology, multimedia, and affective computing. Egon is Editor-in-Chief of Open Computer Science, Area Editor of Pattern Recognition Letters, Section Editor of Journal of Theoretical and Applied Computer Science (JTACS), and Associate Editor of Behaviour & Information Technology. Further, Egon serves as external expert for various agencies (e.g., European Commission, IWT, and ANR), in conference program committees, on boards of advice, and on several journal editorial boards. He frequently serves as invited/keynote speaker, conference chair, and has received several awards, most recently the Journal of the Association for Information Science and Technology (JASIST) 2015 best paper award. Egon guided 70+ students, published 160+ scientific articles, and has several patent applications pending.