Quantifying Sound Quality in Loudspeaker Reproduction

MSc
Computer Science
Game & Media Technology

Utrecht University
Department of Information and Computing Sciences

Kevin van Nieuwenhuizen
3581039

August 25, 2015

Project Supervisors:
dr. J.G. (John) Beerends
dr.dr. E.L. (Egon) van den Broek

Second Examiner:
dr. ir. A.F. (Frank) van der Stappen

Preface

This thesis document is part of my graduation project for a MSc title in Game & Media Technology. It consists of a preface, a technical paper and an appendix. It presents the research of a model that quantifies the sound quality in loudspeaker reproduction. However, this topic is rarely touched upon in Computer Science. This is reflected by the fact that the curriculum of the bachelor and master programs in Computer Science only have a single course related to audio, called "Sound and music technology". This is in contradiction with the essence of Game & Media Technology, as the ability to perceive sound is one of the most important aspects in the field of (traditional) gaming and media. So, I found that it was important to broaden my knowledge on a topic that, while extremely important in the field of gaming and media, I had never touched upon in my study.

One important aspect of audio is the quality of the way it is perceived by the listener. This is determined by two aspects; i) the quality of the signal itself (e.g. codec distortions, recording techniques) , and ii) the quality of the medium used to play the audio. The second aspect is an interesting topic, as there are many different ways to consume audio (e.g. different quality headphones, different quality speakers in different environments). The research I have performed in the last nine months focuses on loudspeakers and has led to a robust model that is a suitable candidate to accurately and objectively quantify the sound quality in loudspeaker reproduction in different environments (e.g. professional recording rooms, average quality livings rooms, low quality hallways), by assessing their acoustic output. It is the first model that works in the acoustic domain that quantifies the sound quality of individual loudspeakers as a whole, and can be used by manufacturers to judge and optimize the quality of loudspeakers for consumers. As a result, it is possible to assess, and indirectly improve, the auditory quality of a portion of gaming and media consumption.

Quantifying Sound Quality in Loudspeaker Reproduction

K. van Nieuwenhuizen*
The Netherlands
J. G. Beerends†
TNO, The Netherlands

ABSTRACT

We present an objective perceptual measurement method for the assessment of the perceived sound quality of loudspeakers, based on the core elements found in the perceptual evaluation models as developed within ITU. Instead of quantifying the loudspeaker system itself, the model quantifies the overall perceived sound quality of loudspeakers by assessing their acoustic output. This approach introduces a major problem. We cannot provide an acoustic reference signal to the subject that can be directly compared to the acoustic degraded loudspeaker output. A solution for this problem is proposed by creating binaural recordings of the reproduced reference signal with a Head and Torso Simulator (HATS), using the best quality loudspeakers available, in the ideal listening spot in the best quality listening environment available. The reproduced reference signal with the highest subjective quality given by subjects is compared to the acoustic degraded loudspeaker output. The model is developed using three large databases that contain binaural recorded music fragments played over very low, to very high quality loudspeakers in very low, to very high quality listening rooms. The average error in percentage between the training ($r = 0.90$) and the validation of the model ($r = 0.85$) is 5.6%, showing the high stability of the model. As such, the model is a suitable candidate to accurately quantify the sound quality in loudspeaker reproduction.

1 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 to assess 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] industry. These models take the signal adaptive properties of the system under test into account by feeding it with real world signals. They measure the quality of the output signals by processing a reference and a degraded signal using a psychoacoustic model, resulting in a representation resembling the internal representation of signals inside our head. The difference between the internal representations is processed by a cognitive model. The result is used to create an objective rating that predicts the subjective rating of the quality of the degraded signal (see Figure 1). This subjective quality is expressed in terms of subjects' Subjective Mean Opinion Score (sMOS), on a scale from 1 to 5 [10] (see Table 1). So, this approach quantifies the quality of the output of a system under test and does not characterize it directly.

Traditional research in loudspeaker reproduction quality follows a classical approach. Instead of characterizing the perceived sound quality produced by loudspeakers, one quantifies the loudspeaker system directly. An extensive overview of this approach is given by Toole [11]. Although a loudspeaker is to a large extent a linear time invariant system, assessing its quality is difficult. This is due to the fact that a one dimensional input music signal (i.e. amplitude 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. Comb filtering and room resonances have a major impact on the reproduction quality.

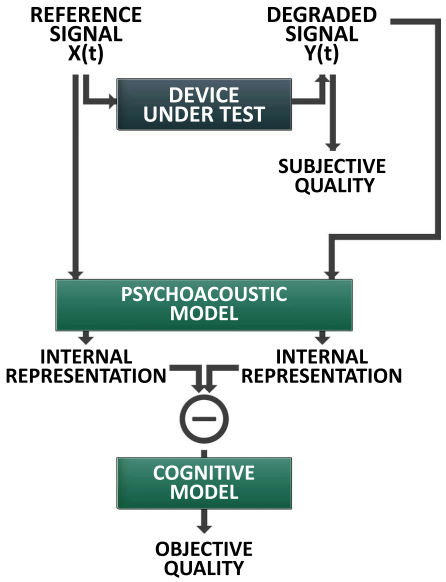

Figure 1: Overview of the basic principle used in models for the perceptual evaluation of audio signals. The psychoacoustic and cognitive model are used to create an objective quality measure that uses the reference signal $X(t)$ and the degraded output, $Y(t)$, of the device under test. This objective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

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 diffuse field. Characterization of how a loudspeaker radiates 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. Thus, when a loudspeaker is reproducing a music signal in a room, all system measurements (e.g. on- and off-axis pressure response, power response, directivity index, harmonic distortion, rub and buzz [12][13][14][15]) will not be able to predict the perceived sound quality. A loudspeaker in a room may sound excellent with one music signal, while the same set up may show low quality with another music signal. Furthermore, additional complications are introduced by our intelligent binaural processing of the input, giving us the possibility to separate sound source

*e-mail: kevinvannieuwenhuizen@gmail.com
†e-mail: john.beerends@tno.nl

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

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).

properties from listening room properties.

Instead of quantifying the system under test, an unconventional approach can be chosen, which quantifies the perceived sound quality produced by loudspeakers using the acoustic output. Tan et al. [16][17] 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 distortions of loudspeakers. Hence, they did not take into account the influence of the listening room and listening position, which both have a dominant impact on the final perceived loudspeaker reproduction quality. Gabrielsson et al. [18][19] 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, spaciousness). However, they did not develop an objective measurement method using this data. Conetta et al. [20][21] used the idea 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 the overall perceived sound quality of loudspeakers 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 the overall perceived sound quality of individual loudspeakers in a wide variety of environments, using a large and diverse data set of music fragments. However, the introduction of this generalization introduces two major problems:

Idealized Reference Signals: In the assessment of an electric input and electric output device, the reference signal used as input of the perceptual model can also be used in the subjective test. 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.

Theoretically, there are two different exact reference approaches possible: i) "here and now", where we have the illusion that the reproduced sound is present in the listening room and ii) "there and then", where we have the illusion that we are present in the room where the recording was made [22]. Both approaches are valid HiFi goals, but require different recording and play back 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 recorded in the anechoic room) with the playback of the anechoic recording. "There and then" requires standard recordings that have to be evaluated using a HATS recording of the "live" event and a HATS recording of the room reproduction. Further, both of these recordings have to be assessed with a correct, individually equalized headphone.

This paper will take a pragmatic approach; binaural recordings of the reproduced signals are made using a HATS, while subjects judge the loudspeaker output on the same listening spot as the recordings. Reference recordings are made using the best quality loudspeakers available, in the ideal listening spot in the best quality environments available. The overall sound quality of the reproduced reference signals are judged by subjects using the sMOS, and the reference recording with the highest sMOS is compared to the acoustic degraded loudspeaker output (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.

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, 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] and is extended with an improved masking algorithm, based on the idea of lateral inhibition [23]. 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 quantification of the loudspeaker system itself or on specific aspects of the acoustic output, the model quantifies the overall perceived sound quality of loudspeakers as a whole.
- Robustness: The model can be used on a wide variety of loudspeakers in a wide variety of listening environments.
- Stability: The model accurately quantifies the sMOS of loudspeaker systems that have not been used in the training of the model.

Section 2 introduces a general overview of the model, as well as the optimization of the model variables. Section 3 presents an overview of the subjective tests used to develop and validate PREQUEL. Results are presented in Section 4. Section 5 presents the conclusions based on this research.

2 THE PERCEPTUAL REPRODUCTION QUALITY EVALUATION FOR LOUDSPEAKERS (PREQUEL)

A general overview of PREQUEL can be found in Figure 2. Each consecutive step performed by the psychoacoustic model is explained in Subsection 2.1 and each consecutive step found in the cognitive model is explained in Subsection 2.2. The model contains a set of variables θ that are optimized using the approach in Section 4. The algorithm used for the optimization of these variables is described in Subsection 2.3. The values that are found in the optimization are used for all the data processed by the model.

2.1 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

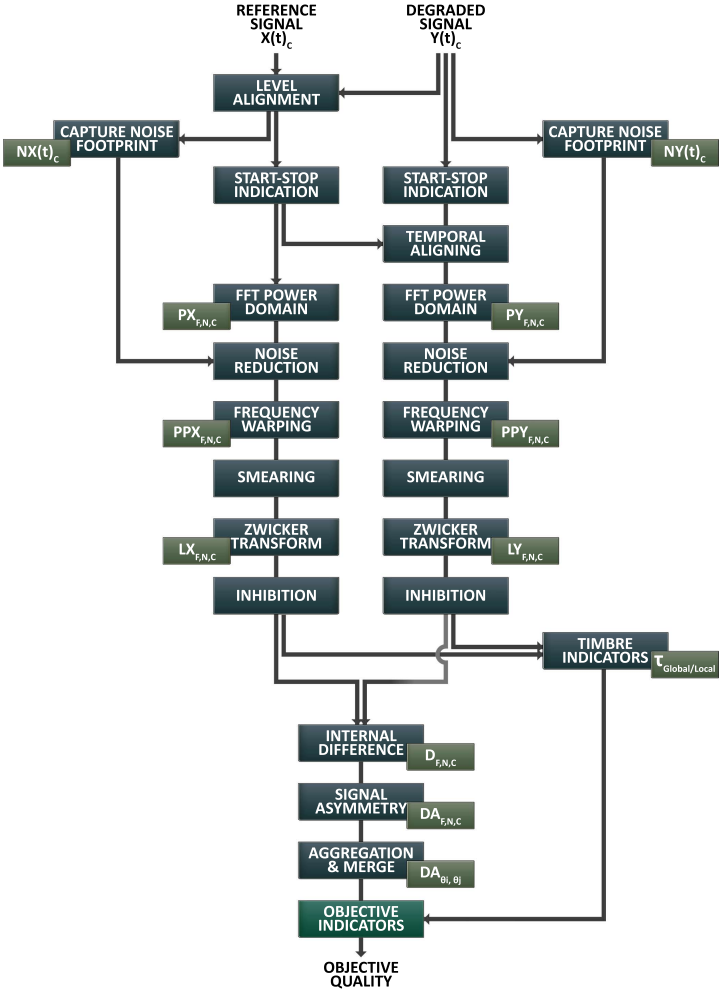

Figure 2: An overview of PREQUEL.

the music fragment starts. Binaural recordings of the reference signal were made with a HATS, using the best quality loudspeakers available, in the ideal listening spot in the best quality environments available. The overall sound quality of the reproduced reference signal is judged by subjects using the sMOS. The reference recording with the highest sMOS is used as the input $X(t)_c$ for the model, where c represents the left or right channel. The degraded signal $Y(t)_c$ 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 by multiplication with a power scaling factor S_p.

The same 40 dB SPL sine wave is used to calibrate the psychoacoustic (Sone) loudness scale using Zwicker's law [24]. Further, the integral of the loudness density, over the Bark frequency scale, is normalized to 1 Sone using a loudness scaling factor S_l.

Level Alignment: The overall power level of the reference signal $X(t)_c$ is scaled to match the overall power level of the degraded signal $Y(t)_c$. The amount of scaling is determined using the ratio of the powers in $X(t)_c$ to $Y(t)_c$.

Start Stop Indication: Recordings of the silent periods at the beginning and end of $X(t)_c$ and $Y(t)_c$ 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 normal 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 $X(t)_c$ and $Y(t)_c$. The samples within this frame are considered noise if their average absolute power is within a range of 0 to 3σ of $\overline{X}$. The value 3σ was found to give the best noise detection.

Further, all consecutive samples at the beginning and end of the signal that are classified as noise are cut from the signal. Thus, after this operation, the reference and degraded signals only contain the music fragment without the silence at the beginning and end of the file. This procedure mimics the behavior of the subjects, which ignore low back ground noise levels in a room when they judge the loudspeaker reproduction quality [25].

Temporal Aligning: 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 Equation 1, where f is $X(t)_c$, g is $Y(t)_c$ and n is the lag coefficient. The overlapping intervals of $X(t)_c$ and $Y(t)_c$ (after applying the lag coefficient to $Y(t)_c$) are used in the remainder of the pipeline.

$$(f \star g)[n] = \sum_{m=-\infty}^{\infty} f^*[m]\, g[m+n] \tag{1}$$

Windowed Fourier Transformation: The human ear performs a time-frequency analysis. Therefore, the algorithm applies a windowed FFT with a Hamming window (see Equation 2) on $X(t)_c$ and $Y(t)_c$, where n is the amplitude per sample of the signal and N is the frame size of 21.34 ms.

$$\omega(n) = 0.54 - 0.46cos(\frac{2\pi n}{N-1}) \tag{2}$$

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 $PX_{f,n,c}$ and $PY_{f,n,c}$ (the power per frequency band f and frame index n for both channels).

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 [25]. 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 $PX_{f,n,c}$ and $PY_{f,n,c}$ respectively.

Frequency Warping: The Bark scale (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. This is implemented by binning consecutive frequency bands of $PX_{f,n,c}$ and $PY_{f,n,c}$, 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 [26]. The resulting signals $PPX_{f,n,c}$

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

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).

and $PPY_{f,n,c}$ are the pitch power densities of the reference and degraded signals.

Frequency & Temporal Smearing: Classical masking is modelled using a bio-mechanical approach in the pitch-power domain. It partially models psychoacoustic masking along the time and frequency axis and quantifies how the power from one time-frequency cell smears towards neighboring time-frequency cells [6]. Smearing is applied in frequency domain on $PPX_{f,n,c}$ and $PPY_{f,n,c}$ using Equation 3, where θ_1 is an optimized constant variable with a value of 0.05.

$$PPX_{f,n,c} = \theta_1 PPX_{f-1,n,c} + PPX_{f,n,c} \tag{3}$$

It is applied in time domain on $PPX_{f,n,c}$ and $PPY_{f,n,c}$ using Equation 4 and 5, where θ_2 and θ_3 are optimized variables with values 0.99 and 1.0 respectively.

$$PPX_{f,n,c} = \phi(f) PPX_{f,n-1,c} + PPX_{f,n,c} \tag{4}$$

$$\phi(f) = \begin{cases} \theta_2, & \text{if } f \geq 500. \\ -\frac{\theta_3-\theta_2}{500} f + \theta_3, & \text{otherwise.} \end{cases} \tag{5}$$

Zwicker Transformation: The reference and degraded pitch power densities are transformed to loudness densities in Sone per Bark using Zwicker's law [24] (see Equation 6), where THR_f is the absolute threshold for the minimum audible field. The Zwicker power, γ, is equal to the optimized variable θ_4 with value 0.145. The function transforms $PPX_{f,n,c}$ and $PPY_{f,n,c}$ to their corresponding loudness densities $LX_{f,n,c}$ and $LY_{f,n,c}$ as functions of time and frequency.

$$LX_{f,n,c} = S_l \left(\frac{10^{THR_f}}{0.5} \right)^{\gamma} \left[0.5 + 0.5 \frac{PPX_{f,n,c}{}^{\gamma}}{10^{THR_f}} - 1 \right] \tag{6}$$

Frequency & Temporal Inhibition: Masking at a bio-mechanical level is implemented using time-frequency smearing. However, masking is also the result of a lateral suppression at a neural level, where firing neurons suppress the firing rate of nearby neurons [23]. This 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 is applied in frequency domain on $LX_{f,n,c}$ and $LY_{f,n,c}$ using Equation 7, where θ_5 has the optimized value of 0.3.

$$LX_{f,n,c} = LX_{f,n,c} - \theta_5 \left(LX_{f-1,n,c} + LX_{f+1,n,c} \right) \tag{7}$$

It is applied in time domain on $LX_{f,n,c}$ and $LY_{f,n,c}$ using Equation 8, where the optimized variable θ_6 has a value of 0.4.

$$LX_{f,n,c} = LX_{f,n,c} - \theta_6 LX_{f,n-1,c} \tag{8}$$

Timbre Indicators: An important aspect of sound quality is the balance between low and high frequencies. This is characterized as its tone color or timbre. If this balance sounds unnatural, subjects will perceive a low sound quality. The ratio of the average loudness between the low (24 Hz - θ_7 Hz) and high (θ_8 Hz - 24,000 hz) frequencies in $LX_{f,n,c}$ and $LY_{f,n,c}$ is calculated, resulting in global timbre values $T1X_c$ and $T1Y_c$. θ_7 and θ_8 are variables with optimized values of 3,400 and 3,000 Hz respectively. The ratio τ_1 (see Equation 9) is later used in the prediction of the sMOS.

$$\tau_1 = MAX \left(\frac{T1X_{left}}{T1Y_{left}}, \frac{T1X_{right}}{T1Y_{right}} \right) \tag{9}$$

A second global timbre indicator is calculated using the ratio of the average loudness between the low (24 Hz - θ_9 Hz) and high (θ_{10} Hz - 24,000 Hz) frequencies in $LX_{f,n,c}$ and $LY_{f,n,c}$. θ_9 and θ_{10} are variables with the optimized value of 1,000 Hz. The results are the timbre values $T2X_c$ and $T2Y_c$. The ratio τ_2 (see Equation 10) is later used in the prediction of the sMOS.

$$\tau_2 = MAX \left(\frac{T2Y_{left}}{T2X_{left}}, \frac{T2Y_{right}}{T2X_{right}} \right) \tag{10}$$

Calculation of the Internal Difference: Two signals that only differ in overall loudness need a minimum difference in order to be discriminated. This is modelled in the form of a self-masking algorithm that uses the raw disturbance density $RD_{f,n,c}$ (see Equation 11).

$$RD_{f,n,c} = ABS \left(LX_{f,n,c} - LY_{f,n,c} \right) \tag{11}$$

The self-masking algorithm is applied using Equation 12 and 13, where θ_{11} and θ_{12} are equal to the optimized values of 0.3 and 0.6 respectively. The algorithm pulls the raw disturbance density towards zero. This represents a dead zone (i.e. before a time-frequency cell is perceived as distorted) and models the process of small time-frequency level differences being inaudible. The result is a disturbance density $D_{f,n,c}$ as a function of time and frequency.

$$D_{f,n,c} = \begin{cases} MAX(0, RD_{f,n,c} - M_{f,n,c}), & \text{if } LY_{f,n,c} > LX_{f,n,c} \\ MAX(0, RD_{f,n,c} - M_{f,n,c}) \theta_{11}, & \text{otherwise.} \end{cases} \tag{12}$$

$$M_{f,n,c} = MAX \left(LX_{f,n,c}, LY_{f,n,c} \right) \theta_{12} \tag{13}$$

2.2 Cognitive Model

Asymmetry: When the system under test introduces a distortion in the input, it will in general result in an output that is clearly composed of two different percepts, the input signal and the introduced distortion. When the distortion is introduced by leaving out a time-frequency component, the resulting output signal cannot be decomposed into two different percepts. This results in a distortion that is less objectionable. This effect is modelled by calculating an asymmetrical disturbance density $DA_{f,n,c}$ using Equation 14, which is applied to the reference and degraded signals $LX_{f,n,c}$ and $LY_{f,n,c}$. $\theta 13$ is a variable that is optimized to a value of 0.1.

$$DA_{f,n,c} = D_{f,n,c} \left(\frac{LY_{f,n,c}}{LX_{f,n,c}} \right)^{\theta_{13}} \tag{14}$$

Aggregation over Time and Frequency: The asymmetrical disturbance density $DA_{f,n,c}$ is integrated along the frequency axis. The result is $DA_{L_i,n,c}$, where L_i is the L_p norm used for the integration, ranging from L_1 to $L10$.

Further, the left and right channel of $DA_{L_i,n,c}$ are merged by calculating the maximum disturbance over left and right in each frame n. The merged disturbance density, $DA_{L_i,n}$, is integrated along the time axis. This results in DA_{L_i,L_j}, where L_j is the L_p norm used for the integration, ranging from L_1 to $L10$. The output of the model is a vector Ω that consists of DA_{L_i,L_j}, τ_1 and τ_2. Ω is used in Section 4 to predict the overall sound quality of loudspeakers using multiple linear regression.

2.3 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 Section 2 and 16 variables that are needed to prevent instabilities of the model. Each variable is given a lower and an upper bound (based on the existing perceptual evaluation models as developed within ITU for speech [1][2][3][4][5] and music [6][7][8]), and a Δ value defined by the user that describes a finite increment of the variable. These values are used in the optimization algorithm.

One approach for optimization is to use a brute force algorithm to find the best solution. 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. This makes it infeasible to calculate with higher values of M, due to hardware restrictions. 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, was implemented. While this algorithm does not guarantee to find the optimal solution, it is a lightweight optimization strategy that provided 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 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 a user defined number of 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.

3 SUBJECTIVE EXPERIMENTS

Three different experiments were run for the training and validation of PREQUEL. Each experiment used a sequence of 6 music fragments, which were chosen on the basis of their high quality as judged by expert listeners. A total of 12 musical fragments were used that included classical large orchestras, opera/choir, solo instruments and pop/rock recordings. The six fragments in each experiment had a duration of about 30 seconds and were played consecutively with silences of about four seconds between each fragment. Each fragment was individually level aligned for the optimal play back level relative to the other fragments. The loudest fragment (rock) was played at a level of about 90 dB (A), fast averaging. The softest fragment (solo harpsichord) was played at a level of about 65 dB (A), fast averaging. All fragments in each experiment were binaurally recorded using a HATS.

Each experiment was performed by six subjects, ranging from naive and trained listeners, to expert listeners. A total of 18 subjects were used, consisting of 16 males and 2 females, with an age ranging from 22 to 74. 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. A 10 point evaluation scale, based on the school reporting system used in The Netherlands, was used for the judgments, where 1 stands for "bad" and 10 stands for "excellent". This scale was chosen because it provided the most natural opinion scale to express a quality opinion in The Netherlands. Each subject had a training session before each experiment started, by providing them a direct comparison of music fragments played over any system they would like to hear.

The first two experiments were performed using two high quality professional listening rooms with excellent acoustic properties. Subjects were seated at 3 to 4 different positions in each room and judged the quality of 9 loudspeaker systems, ranging from high quality standard studio monitors (electro dynamic and electro static) and high quality surround radiating systems, to average consumer type systems and very low quality PC loudspeakers. 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 (e.g. standard living room, hallway, kitchen). Subjects were seated at 2 different positions and judged the quality of 8 loudspeaker systems, ranging from high quality surround radiating systems (electro dynamic), to normal consumer type systems and very low quality loudspeakers. Furthermore, the experiment included a 4 channel surround system and a number of 2 channel room reverberation algorithms. There were a total of 22 different loudspeaker reproduction evaluation setups, which resulted in a total of 132 fragments that had to be judged.

Consistency checks of each experiment show that all subjects have a correlation of at least 0.77 (the worst naive subject) between their private opinion and the average opinion of the group, while the best subject (the best expert) 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, room resonances), the consistency in judgement is very high, verifying the high relevance of the subjective data when describing the overall perceived sound quality.

The subjective scores given by the subjects are standardized per experiment using Equation 15, where Z is the standardized subjective score, X the original subjective score, μ the mean and σ the standard deviation of the data of each subject.

$$Z = \frac{X - \mu}{\sigma} \tag{15}$$

The sMOS for each fragment is the mean of its corresponding standardized subjective scores. Further, the sMOS of each fragment was normalized for each experiment individually to the ITU five point scale (see Table 1) to adjust the results presented in this paper (see Section 4) to a widely accepted standard scale. The normalization was performed using Equation 16, where Z is the standardized sMOS, L the lowest standardized sMOS and H the highest standardized sMOS in each experiment. The result is the normalized sMOS.

$$\textit{Normalized sMOS} = \frac{4(Z - L)}{H - L} + 1 \tag{16}$$

4 RESULTS

The three experiments described in Section 3 are used to create three loudspeaker databases. A database consists of a collection of binaural recorded music signals, with corresponding index k, and the normalized sMOS per signal, sMOS_k. These signals are used in the model as the degraded recording $Y(t)_c$. The signal with the highest sMOS_k in the database that uses the same music fragment as $Y(t)_c$ is used as the reference recording $X(t)_c$. The data from the first two experiments is used to create database DB_1 and DB_2. The data from the third experiment is used to create DB_3. Furthermore, each database is split in two parts by dividing each database in four equal intervals based on the sMOS, and splitting each interval in two equal parts. This method enforces the full range of sMOS in each new database, making sure that they remain as balanced as the originals. This results in two sets, A and B, that each contains one half of the databases.

The performance of the model is measured in terms of the correlation coefficient r between the vector Ω of each fragment and its corresponding sMOS using multiple polynomial regression. The final results are expressed as the mean of the prediction of each music signal per loudspeaker in each room. The following two paragraphs will explain the training and validation of the model.

Training: The training set A is used to develop a robust model that is trained context independently on the data in A in such a way that it is able to quantify the subjective sound quality of loudspeaker systems. First, all model variables are trained (see Subsection 2.3) using A, resulting in the optimized variables θ. Next, each signal in the training set is processed by the model using the optimized variables, resulting in Ω_k per signal.

Further, multiple polynomial regression is used to model the relationship between a vector of predictor variables $\omega_k \subset \Omega_k$ and the sMOS_k. This is done by transforming ω_k to a single predictor ψ_k using a nonlinear function, and fitting a monotonic polynomial P through the data points (ψ_k, sMOS_k). The nonlinear function that is optimized in the regression can be found in Equation 17. A 3^{rd} order monotonic polynomial is used for the fit as it provided the best results while maintaining monotonicity.

$$\psi = -1\left[DA_{L_1,L_1}\tau_1\right] - 0.045\left[DA^2_{L_2,L_2} + \tau_1\tau_2\right] + 0.034\left[\tau_2 + \tau_2^{-1}\right] \quad (17)$$

However, this approach introduces a problem due to the robust nature of the model. 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. Though, due to the context of each experiment, the same degradation in each database may result in a different sMOS due to voting preferences or the balance of conditions in a test. 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. Further, this signal is presented in two different experiments. The signal has the highest quality in the context of one experiment and the lowest quality in the context of the second experiment. This results in two different sMOS given by subjects to the same signal due to the difference in context of each experiment.

The solution to this problem is to perform the multiple polynomial regression using 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. The result is a robust model that can predict the sMOS context independently. The resulting r values of the multiple polynomial regression are 0.94, 0.90 and 0.88 for $DB_1 \subset A$, $DB_2 \subset A$, and $DB_3 \subset A$ respectively.

Validation: The validation of the model is performed using a blind prediction of the sMOS on the signals in the validation set B. Each signal is processed by the model using the optimized variables θ found in the training, resulting in an Ω_k per signal. Equation 17 is used to transform $\omega_k \subset \Omega_k$ to the single predictor ψ_k. Further, P_m is used to transform ψ_k to the prediction of the sMOS. The results can be found in Figures 3, 4 and 5. The Figures show the 95% confidence interval and the ideal linear regression $Y = X$. The resulting r values between the prediction and the sMOS for $DB_1 \subset B$, $DB_2 \subset B$ and $DB_3 \subset B$ are 0.86, 0.92 and 0.78 respectively.

The stability of the model is validated by comparing the r values of the training with the r values of the validation. The error in percentages between these values is 8.5%, 2.2% and 11.0% respectively. The results for DB_3 are slightly less stable than the other two databases. This could be due to the higher difficulty of taking into account all possible degradation parameters in the third experiment. This is reflected by the fact that subjects had a harder time to form a coherent opinion with other subjects (lowest $r = 0.50$), as well as with the average opinion of the group (lowest $r = 0.77$). These values are significantly higher in the first and second experiment, where the lowest correlation between two subjects' opinion is 0.65 for both experiments and the lowest correlation between a subjects' opinion and the average opinion of the group is 0.86 and 0.84 respectively. Thus, it is expected that the stability of the model decreases as subjects have a harder time predicting the subjective quality. Nonetheless, the results show high stability of the model when quantifying the subjective sound quality of loudspeaker systems that produce the same type of distortions introduced in the training set A.

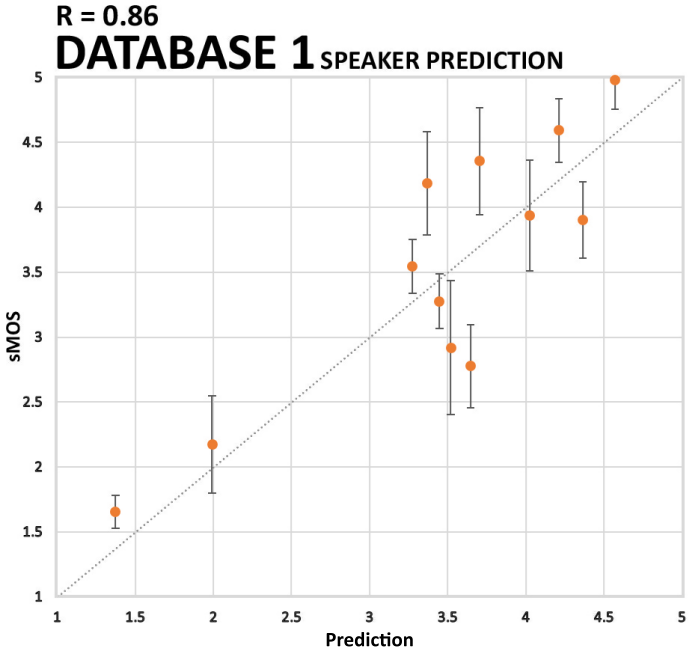

Figure 3: The blind prediction for database $DB_1 \subset B$, plotted per loudspeaker. The graph shows the 95% confidence interval and the ideal linear regression $Y = X$. The correlation coefficient between the predictor values and the sMOS is 0.86.

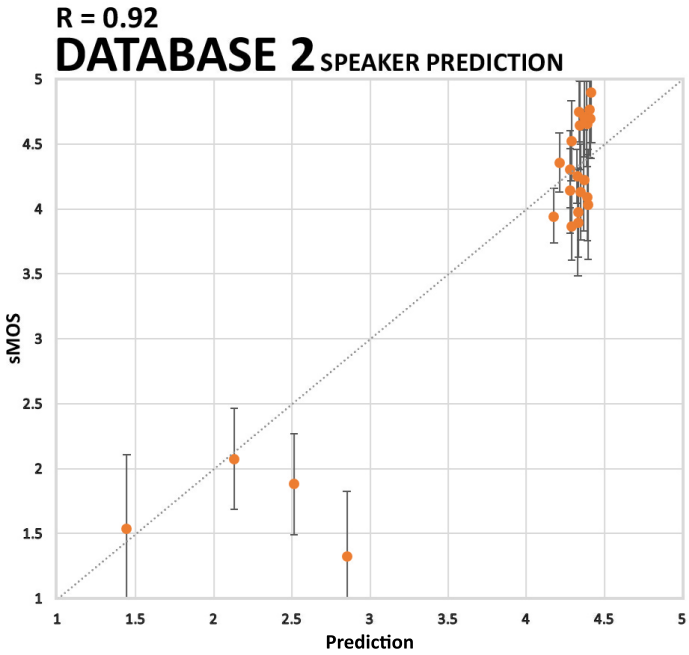

Figure 4: The blind prediction for database $DB_2 \subset B$, plotted per loudspeaker. The graph shows the 95% confidence interval and the ideal linear regression $Y = X$. The correlation coefficient between the predictor values and the sMOS is 0.92.

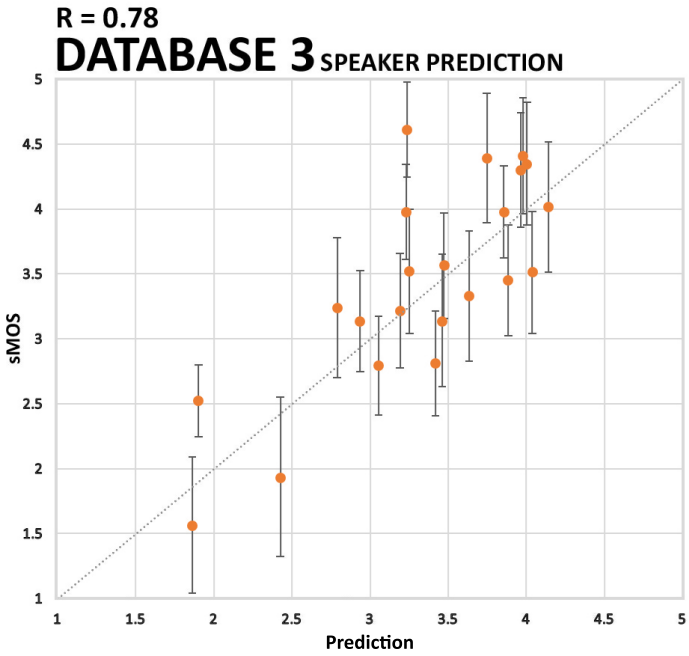

Figure 5: The blind prediction for database $DB_3 \subset B$, plotted per loudspeaker. The graph shows the 95% confidence interval and the ideal linear regression $Y = X$. The correlation coefficient between the predictor values and the sMOS is 0.78.

5 DISCUSSION

This paper presents a unique loudspeaker reproduction quality measurement model baptized the Perceptual Reproduction Quality Evaluation for Loudspeakers (PREQUEL). Whereas previous research focused on the quantification of the loudspeaker system itself or on specific aspects of the acoustic output of loudspeakers, this paper focuses on the overall perceived sound quality of individual loudspeakers in a wide variety of listening environments, using a large and diverse data set of music fragments.

This perceptual measurement approach introduces two major problems. The first one is that we cannot provide the subject with an acoustic reference that can be directly compared to the degraded loudspeaker output. A solution for this problem is proposed by creating binaural recordings of the reference signal using the best quality loudspeakers available, in the ideal listening spot in the best quality listening environments available. The recorded reference signal with the highest subjective quality given by subjects is compared to the acoustic degraded loudspeaker output. The second problem is the presence of different levels of background noise in the listening environments. This is solved by introducing a noise suppression algorithm that operates on both the recorded reference and degraded signals.

The solutions to these problems led to the development of a model that uses an objective quality measurement to predict the sMOS of subjects that were asked to judge the overall quality of a set of loudspeaker reproduction systems. Consistency checks performed on the subjective data show a correlation of at least 0.77 between subjects' private opinion and the average opinion of the group, verifying the high relevance of the subjective data when describing the overall perceived sound quality.

Three databases, based on the gathered subjective data, were created for the training and validation of PREQUEL. The model is trained in such a way that it is able to accurately predict the sMOS context independently. Results from the validation show that the model is stable and is a suitable candidate to accurately quantify the sound quality of individual loudspeakers, in a wide variety of settings, based on the distortions introduced in the training phase.

However, the model is also able to quantify the sound quality of individual fragments. The average correlation coefficient in the training phase when plotted per individual fragment is 0.86. The average correlation coefficient in the validation phase when plotted per individual fragment is 0.79. The error in percentage between the average correlation of the training and validation is 8.1%, which is comparable to the average error when plotted per loudspeaker (7.2%). Thus, the model assess the quality of a single fragment recorded using a certain loudspeaker in a certain environment with an accuracy comparable to the prediction per loudspeaker.

It is important that further model validations are carried out by taking into account the combined effect of loudspeaker distortions, as well as other types of distortions (e.g. amplitude clipping, low bit rate audio coding, time clipping). It is expected that the performance of the model will drop when the model is validated on these type of distortions that are unknown to the model. However it is also expected that a retraining will allow the model to cope with these new distortions.

Furthermore, it is important to investigate the influence of different HATS on the performance of the model. 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. It is interesting to observe to what extent each of these parameters will influence the results of the model. For instance, it is expected that the usage of two different HATS to record the reproduced reference signals and the degraded acoustic output respectively results in significantly lower correlations of the model, if the HATS have different types of ears. We expect that the current model needs to be retrained slightly when dealing with these types of variations, to cope with the differences in Head Related Transfer Functions.

Finally, the optimization strategy used in this paper does not guarantee an optimal solution. Instead, it produces a local optimum that is presumed to be good enough. This optimization strategy is used due to hardware restrictions. A simple solution is to improve the capabilities of the hardware using vertical or horizontal scaling. Scaling vertically results in an upgrade of hardware, which improves the computing power. Scaling horizontally introduces more hardware, resulting in more efficient parallelization. Sufficient scaling may lead to the feasibility of a brute force approach in reasonable time. Furthermore, variations of our current algorithm have the possibility to perform better, depending on the situation. An example would be to introduce a form of relaxation that accepts solutions that are worse than the current optimum found in the algorithm. This allows for a more extensive search in the solution space.

6 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.

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. *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 Pocta and John G Beerends. Subjective and objective assessment of perceived audio quality of current digital audio broadcasting systems and webcasting applications. 2015.

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

[11] Floyd E Toole. *Sound reproduction: loudspeakers and rooms*. Taylor & Francis, 2008.

[12] Wolfgang Klippel and Ulf Seidel. Measurement of impulsive distortion, rub and buzz and other disturbances. In *Audio Engineering Society Convention 114*. Audio Engineering Society, 2003.

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

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

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

[16] 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.

[17] 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.

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

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

[20] 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.

[21] 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.

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

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

[24] Eberhard Zwicker and Richard Feldtkeller. *Das Ohr als Nachrichtenempfänger*. Hirzel, 1967.

[25] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

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

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

Appendix

1 Introduction

The appendix contains information that cannot be found in the paper. Section 2 gives an overview of the content of the deliverables for this research. Section 3 gives instructions on how to use the application of the model and gives an in-depth overview of the time and memory complexity of the algorithms used in the model. Section 4 gives an overview of the materials used in the experiments.

2 Deliverables

The following list contains the structure of the folders of the deliverables. This list, called "Readme.pdf", can also be found in the root folder. The root folder has the following structure:

Model:

Config Data:

* Contains the Config.xml file needed to run the application.

Databases:

* Contains two databases that can be used to run the application.

Documentation:

* Contains the documentation website for the code.
  Click on "index.html" to open the website.

Project:

* Contains the Visual Studio 2013 project for the application.
* Contains a "Readme.pdf" for the application.

Build:

* Contains an executable of the application.
* Contains an ".cmd" file used to start the application.
  Do not forget to change the path of the argument in this file.

Documents:

Results:

* Contains the correlation coefficients of all experiments.
* Contains the results of the training & validation phase.

Paper:

* Contains the paper bundled with the preface and the appendix.

3 Model

This Section is divided in two Subsections. Subsection 3.1 presents an overview of how the application works. Subsection 3.2 presents an overview of the time and memory complexity of the application.

3.1 Overview

The executable of the application accepts a single argument. The argument is the absolute path to the configuration ".xml" file that is needed to run the application. An example of this file can be found in the "Model\Config Data" folder. It is important to change this file, as the application uses absolute paths to find certain files and folders.

The model uses the databases described in the configuration file as the input of the model. A database is described with the following information. The name of the database should be a unique identifier. This name is used as the identifier of the database in the application and when writing the results to disk. There is also a field with the size of the database. It is important that this number is correct, because it is used to validate the cache. The path to the database points to a folder with all the reference and degraded signals in this database. The audio data path is the path to the cached ".xml" signal files that are written to disk. If there are no cached ".xml" files, the application will create these files the first time it is run. Note that this feature can be turned off at the top of the configuration file. The documentation path is used for the documentation about the database. Examples of these files can be found in the "Model\Databases\Subset Database X_Documentation" folder. This file is responsible for all information regarding the signal in this database (e.g. sMOS, CI95, sample rate, relative reference & degraded paths).

The predictors describe the function that produces ψ_k (see Section 4 of the paper) and the polynomial P. It is important to note that the configuration file only accepts a single polynomial that is used for all input. It is recommended to leave this polynomial as default and apply the polynomials for each database context dependently using a statistical program (e.g. excel).

The variables are categorized per component in the pipeline. You can refer to the paper when needing to change a parameter of a certain component. The actual purpose of the variables can be found in the code.

The optional part of the configuration file is only needed when optimizing. Thus, if one wants to use the model as it is supposed to (i.e. processing signals), this can be deleted. However, this should be filled in when optimizing. The information needed is straight forward. There is room for expansion for different optimization algorithms. The configuration file, and the corresponding parser in the code, is built in such a way that it is easy to add or delete information. The same can be done when implementing a polynomial per database instead of how it is currently.

The application can be run when everything is filled in correctly. It is a simple console application that displays all the necessary information in the console window. Note that the application should be compiled in x64, if the Visual Studio 2013 project fails to compile with the default settings. The results of the application can be found in the specified results folder after the application is done with the calculations. The results consist of a log file and the results per database, categorized per folder. When performing

the optimization, the results are bundled based on the hill climbing instance in the code that produced the results.

3.2 Complexity

The complexity of the application is defined as the time and memory the application consumes when executing the pipeline and its components (see Figure 2 of the paper) for a single output vector Ω. It does not include the optimization of the variables (see Subsection 2.4 of the paper) and the multiple polynomial regression. The multiple polynomial regression is performed by an external program called "Uni Huge", which is a statistical program developed within TNO.

Time Complexity: The pipeline can be divided in two categories; i) all operations before the Fast Fourier Transformation (FFT), and ii) the FFT and all operations after the FFT. The first category performs all operations on 1 dimensional signals (a function of time). The second category performs all operations on two dimensional signals (a function of time and frequency). The number of signals and channels processed by the pipeline can be regarded as constant values, because they are both fixed at a value of 2.

Almost all components in the first category have a time complexity that is linear in the number of samples N per signal. However, the calculation of the cross correlation is more complicated. It is calculated by reduction to the convolution of $f^*(-t) * g$ (see Equation 1), where f is $X(t)_c$ and g is $Y(t)_c$.

$$f \star g = f^*(-t) * g \tag{1}$$

One of the fastest ways to calculate this, is by calculating the circular convolution of two signals using the FFT of each signal (padded with zero values at the end of the signals), multiplying them pointwise, and then performing the inverse FFT. This algorithm has a time complexity of $\mathcal{O}(NLogN)$.

The calculation of the FFT in the second category can be done efficiently with a time complexity of $\mathcal{O}(MLogM)$ if the length of the signal used for the FFT, M, is a power of two. Further, the pipeline uses windowed FFT, which divides the signal in $\frac{N}{M}$ segments based on a window, where M is the size of the window, and performs a FFT on these segments. Thus, the time complexity to calculate the FFT on a signal is $\mathcal{O}(\frac{NM}{M}LogM)$. This can be simplified to $\mathcal{O}(NLogM)$. The result of the FFT is a signal as a function of the time frames n and frequency bands f, where n is equal to $\frac{N}{M}$ and f is equal to $\frac{M}{2}-1$. The rest of the components in this category have a time complexity based on f and n, $\mathcal{O}(\frac{N}{M}\left(\frac{M}{2}-1\right))$. This can be reduced to $\mathcal{O}(\frac{NM}{M})$, resulting in $\mathcal{O}(N)$.

The upper bound of the whole pipeline can be written as $\mathcal{O}(NLogN)$ if $N \geq M$. This is always true, due to the mandatory 1 second of silence at the beginning of each signal. Further, each signal is sampled at 48 kHz, resulting in a signal of at least 48.000 samples. However, this upper bound is only valid when reading the audio files from disk, due to the cache the application uses. The cache saves the state of all signals processed by the application after the frequency warping, and writes it to disk. These files are loaded in memory and any subsequent calculations for these signals will use the cache. Thus, the components in the first category are not calculated in this approach, resulting in a time complexity of $\mathcal{O}(NLogM)$.

The time complexity of the application changes to $\mathcal{O}(DNLogM)$ when processing multiple signals, where D is the number of signals that need to be processed. The increase of the factor D is compensated by the fact that each vector Ω is calculated on a separate thread, resulting in a significant increase of speed depending on the hardware used.

Memory Complexity: The cache has no upper bound on its size. Thus, all signals that are processed by the application are loaded in memory. This is important when optimizing the model variables, because the I/O per optimization round is a significant bottleneck. Thus, the cache results in a significant increase in performance, at the cost of a larger memory footprint. The memory complexity depends on the number of signals D processed by the application and the size S in memory of each fragment. Thus, the memory complexity of the algorithm is $\mathcal{O}(DS)$.

Each signal used in the application is in stereo and is sampled at 48 kHz. Each sample in a signal has a size of 16 bits and has a duration of about 30 seconds, with at least 1 second of silence recorded before the music fragment starts. This results in a size of approximately 6 MB per signal. The input of the model consists of a reference and degraded signal, resulting in an input of size 12 MB. A total of 348 pairs were processed by the model, resulting in a memory footprint of 4 GB. This could be a problem, due to the fact that this footprint grows linearly with the number of signals processed by the application. A previous iteration of the application had a cache that loaded blocks of information in memory, resulting in a maximum memory footprint that could be controlled by the user. Thus, a solution to a potential memory problem is to define an upper bound on the size of the cache by swapping blocks of information in and out of memory.

4 Experiments

This section gives an overview of the materials used in the experiments. This includes the different loudspeaker systems and the music fragments used in the experiments. The fragments that are used in the first experiment were used in the development of the MPEG standard. The corresponding MPEG database did not save the original names of these fragments. Instead, they are described by the type of music signal (e.g. accordion, trumpet). The loudspeakers and the fragments used for each experiment can be found in Figures 1, 2 and 3.

EXPERIMENT 1

Loudspeakers	Fragments
Bloomline (4x dynamic)	MPEG Bizet
KEF Cresta (2x dynamic)	MPEG Trumpet
Home build (2x dynamic)	MPEG Chapman
Aldi Brand X (2x dynamic)	MPEG Accordion
Grimm Audio LS1 (2x dynamic)	MPEG BassGuitar
PC mini speaker (2x active dynamic)	MPEG Percussion

Figure 1: The materials used for the first experiment. Each fragment was played over all the loudspeaker systems in a high quality listening environment with excellent acoustics.

EXPERIMENT 2

Loudspeakers	Fragments
Bloomline (4x dynamic)	Daniel Cross - The Spinroom
KEF Cresta (2x dynamic)	Stanislav Moryto - Per uno solo
Quad ESL-989 (2x electrostatic)	Antonín Dvorák - Slavonic Dance Op. 72
PC mini speaker (2x active dynamic)	William Walton - Set me as a seal upon thine heart
Tannoy Small Studio Monitor (2x dynamic)	Johann Sebastian Bach - Sarabande uit Partita II, BWV 826
Tannoy Large Studio Monitor (2x dynamic)	Georg Philipp Telemann - Ach, Herr, Straf Mich Nicht In Deinem Zorn

Figure 2: The materials used for the second experiment. Each fragment was played over all the loudspeaker systems in a high quality listening environment with excellent acoustics.

EXPERIMENT 3

Loudspeakers	Fragments
Bloomline (4x dynamic)	Daniel Cross - The Spinroom
KEF Cresta (2x dynamic)	Stanislav Moryto - Per uno solo
Samsung TV (2x dynamic)	Antonín Dvorák - Slavonic Dance Op. 72
Cheap Brand X (2x dynamic)	William Walton - Set me as a seal upon thine heart
Bower & Wilkens DM6 (2x dynamic)	Johann Sebastian Bach - Sarabande uit Partita II, BWV 826
Tetra home build + Aldi diffuse field filler 4 channel (4x dynamic)	Georg Philipp Telemann - Ach, Herr, Straf Mich Nicht In Deinem Zorn
Tetra home build + Aldi diffuse 4 channel + Yamaha room simulator Hall in Vienna (4x dynamic)	

Figure 3: The materials used for the third experiment. Each fragment was played over all the loudspeaker systems in three low, to average quality listening environments (i.e. living room, kitchen, hallway).