Audio Engineering Society

Convention Paper

Presented at the 117th Convention
2004 October 28–31 San Francisco, CA, USA

Evaluation of Different Loudness Models with Music and Speech Material

Esben Skovenborg[1,2] and Søren H. Nielsen[2]

[1] University of Aarhus, Dept of Computer Science, Åbogade 34, DK-8200 Århus, Denmark
esben@skovenborg.dk

[2] TC Electronic A/S, Research Department, Sindalsvej 34, DK-8240 Risskov, Denmark
shn@tcelectronic.com

ABSTRACT

The evaluation of twelve models of loudness perception is presented. One of the loudness models is based on a novel algorithm, and another is based on a combination of two known measurement techniques. The remaining models are all implementations of common or standardized loudness algorithms. The ability of each model to predict or measure the subjective loudness of speech and music segments is evaluated. The reference loudness is derived from two listening experiments using the speech and music segments as stimuli. Different statistical measures are employed in the evaluation of the models, so that both the absolute performance of the models and the performance relative to the between-listener disagreement are measured.

1 INTRODUCTION

Most everyday listening situations involve reproduced sound which is a combination of music and speech. The *loudness* of the sound, perceived by the listener, will not only depend on his or her volume setting, but also on the particular source and format of the audio material. Normally the programme material that is broadcast on radio/TV or distributed on CD has been dynamically and spectrally processed, in order to fulfill aesthetical and technical requirements. Such processing may also affect the perceived loudness of the material. Listeners often experience undesirable 'jumps' in loudness between different sources or broadcast channels, and also between different programme segments within the *same* channel.

If the perceived loudness could be *predicted*, it would be possible to measure and control the loudness of the programme material. Loudness perception has been researched extensively in classical psychoacoustics, under laboratory conditions, traditionally using stationary and/or synthetic signals. Moreover, so-called objective loudness measurement procedures have been under continual development for decades. Nevertheless, no loudness model has yet been established as an accurate measure of the loudness of audio material such

as the music, speech, and commercials being broadcast on radio and TV.

The **applications** for such a loudness model would include *loudness meters* [1, 2, 3, 4, 5, 6], as well as devices for *loudness control* [7, 8, 9, 10, 11, 12]. In connection with productions for broadcast, applications of a loudness meter has been studied in connection with monitoring and leveling practices [13, 14, 15, 16, 17]. The "loudness maximizing", achieved by aggressive application of dynamics compression, is frequently performed in the mastering of audio CDs. This practice may lead to an undesirable sound, as well as technical problems [18, 19, 20]. An accurate loudness model may also be desirable for other applications, for example in *sound quality research* where the a set of stimuli might need to be "loudness equalized" so that other (subjective) factors can be investigated independently of loudness [21, 22].

The perceived loudness of music and speech can be accurately measured by means of a controlled listening experiment [23, 24]. The loudness of homogeneous sound segments, with a duration of say 10-15 seconds, may be compared as the overall loudness of each segment is perceived to be fairly constant. The property of the sound which is assessed is its **long-term loudness**. In some applications of a loudness model, it would be desirable to employ a *short-term* loudness measurement in combination with the long-term loudness. However, as it might be difficult to assess the perceived short-term loudness of music and speech in experiments, only the long-term loudness is examined in this paper. The specific relation between the short- and long-term loudness has not yet been established, apart from the notion that the short-term loudness must somehow converge towards the long-term loudness when the analysis window is increased in duration (e.g. [25]).

This paper presents the evaluation of twelve models of loudness perception. The ability of each model to predict or measure the perceived loudness of speech and music segments is tested. Two independent listening experiments were conducted, with a collection of speech and music segments as stimuli. The results of the experiments provide the subjective reference data against which the models are evaluated. Loudness is subjective by nature, and as such it is more difficult to 'measure' than physical properties. In this paper, we propose both a visual method and a statistical method of evaluating the loudness models, while taking both the uncertainty and subjectivity of the reference data into account. Other aspects, important for the application of a loudness model, such as real-time operation, reference level, and metering characteristics, are beyond the scope of this paper.

Most of the established loudness models were originally constructed to measure the loudness of stationary signals, such as noises. We present two new models of loudness that were developed specifically for estimating the perceived loudness of music and speech segments.

1.1 Some aspects of loudness perception

Loudness perception has been studied for centuries and many findings have been described in the literature, e.g. [26, 27]. A few essential properties of the auditory system, related to loudness perception, are summarized below: integration along a perceptual frequency axis, masking, adaptation, compression and integration along the time axis.

1.1.1 Intensity perception

One basic aspect is how we perceive sound with different intensities. The transfer function from the physical magnitude of the sound stimulus to the perceived magnitude is not linear. A good (but not the only) approximation to this transfer function is expressed in *Stevens' power law* [28] (after [29, 30]), in which a characteristic power function is applied to the physical magnitude (intensity), I, to get the perceived magnitude (sensation), S :

$$S = K I^n \tag{1}$$

In eq. 1, n is an empirically derived power corresponding to the particular type of stimulus, and K is a constant adapted to the unit used to make the judgment. For loudness perception, the power, n, is around 0.3 [31]. This applies to loudness well above the threshold in quiet. For low frequencies the power is higher, meaning that a given change in physical intensity will be perceived as a larger change than if a similar intensity change was made at medium or high frequencies. The power law can be reformulated using logarithms:

$$\log S = n \log I + \log K \tag{2}$$

This correspondence between physical and perceived magnitudes is one of the motivations for the *decibel*

scale. For example: With two sounds, one of which is 10 times as intense as the other in the physical domain, the ratio in the sensation domain is: $10^{0.3} \cong 2.0$. Thus, by increasing the intensity of the sound stimulus by a factor of 10 the perceived loudness is only twice as high; so a *compression* takes place. (Other senses have other powers, *n*, some of which are larger than 1, in which case the power law becomes an expansion function [29].)

The perceived loudness can be expressed in *sone* or *phon* (e.g. [27]). The sone scale is a ratio scale, so the value expressed in sone is directly proportional to the perceived loudness. For instance, if one sound has a loudness of 1 sone, another sound, which is perceived as twice as loud, would have a loudness of 2 sone. It may be more convenient and/or accurate to express loudness as a *loudness level*, in phon. The phon scale is similar to the dB scale and the two coincide, for 1 kHz tones, at levels above approximately 40 dB sound pressure level.

When comparing the loudness of two sounds, by adjusting the relative level to match their loudness, the non-linear characteristic in itself is less important. The loudness difference, in this case, is typically expressed in phon or dB.

1.1.2 Spectral aspects

The loudness function expresses the correspondence between the physical level of some audio signal and its loudness. A strong frequency-dependence of the loudness functions exists – often expressed in the *equal loudness contours*, e.g. [32]. Alternatively, the constants *n* and *K* in eq. 1 and eq. 2 could be considered as functions of frequency. An important property of loudness perception, and one which might be surprising at first, is the *spectral loudness summation*. In short, the loudness of a signal increases with its bandwidth, for a constant total signal power. The effect of this property, in the extreme case of the difference between a pure tone and wide band noise (uniform excitation noise or white noise), is 11-18 dB, depending on the SPL [27](sect. 8.3).

The spectral loudness summation is associated with the concept of *critical bandwidth* [33, 27]. Within one critical band, which is approximately 1/3 octave wide at medium and high frequencies, the loudness of a signal, at a given level, is independent of its bandwidth. When the signal's bandwidth grows outside the critical band the loudness increases, even though the total level is kept constant. The 'frequency' axis, along which the loudness integration takes place, is specified in different units, depending on the particular definition of the critical bandwidth; the most common units are the Bark and ERB (equivalent rectangular bandwidth) [31].

Signals of different frequencies also interact in a complex way known as *spreading*, which is closely related to *masking*. An effect of this phenomenon is that not only the actual frequencies of a signal contribute to the loudness, but also to some extent the higher and lower neighboring frequencies. As a consequence of spreading, the loudness functions for pure tones and noise are different. For a 1 kHz pure tone, the dB SPL and phon scales coincide above 40 dB SPL, whereas a noise signal will deviate, partly due to level dependent spreading functions [27, 31].

1.1.3 Temporal aspects

The hearing does not react instantaneously to sound – building up the loudness takes some time. Within certain limits, the loudness is proportional to the energy of a tone or noise burst. This principle is valid up to a duration of about 100 ms, where the loudness of the burst is perceived as equal to a continuous signal of the same amplitude [34, 35]. Furthermore, when a sound ends, the loudness perception does not die out instantaneously. This is related to the *post-masking* effect. Unfortunately, no simple time constant can be specified for the temporal loudness integration. The loudness function depends both on the bandwidth (white noise, narrow-band noise, sine wave) and on the duration of a burst [36, 37, 38].

A more long-term temporal effect is the *temporal threshold shift* that occurs after listening to loud sound for a certain duration, e.g. at a rock concert. The hearing ability is reduced for some time afterwards, but will (hopefully) return to its normal state after some hours.

1.1.4 Spatial aspects

The direction dependent filtering, caused primarily by the outer ear, influences the loudness perception. To complicate matters further, we normally listen with both ears at the same time, leading to the non-trivial phenomenon of *binaural loudness summation* [27]. When listening to a sound in a room the time domain signal is smeared by the room's reverberation characteristics. This filtering is quite complex and furthermore position dependent.

1.1.5 Test signals as stimuli

In order to obtain an understanding of the basic auditory mechanisms psychoacousticians have generally employed audio signals with certain analytic properties. Signals such as constant or pulsating sine waves, noise of varying bandwidth, and tone complexes, have typically been used as stimulus [27, 26]. Although such signals are good for the purpose for which they are used, they are unfortunately quite far from the typical programme material on radio/TV broadcasts, in terms of spectral and dynamic content. Real-world signals, such as music, are typically broadband and fluctuating. Even though much is known about the auditory mechanisms, the highly complex nature of real-world signals presents a challenge to traditional psychoacoustic models.

The research driven by the 'users', such as broadcasters and audio equipment manufacturers, tends to focus on real-world material in what could be called a *top-down approach* – in this case to loudness perception. This is perhaps due to the difficulty of extrapolating results derived from simple signals to very complex ones. The research driven by academic institutions, on the other hand, tends to focus on simple sounds, either in a top-down or bottom-up approach to understand the individual auditory mechanisms, often in isolation.

1.2 The ITU SRG-3 perspective

In September 2000, the WP-6P of the ITU-R set up a Special Rapporteur Group, the *SRG-3*, to investigate into "*Audio metering characteristics suitable for use in digital sound production*" [39]. Specifically, the ITU-R considered "*that listeners desire audio programmes to be uniform in subjective sound level*", "*that current knowledge of human psychoacoustics may make it possible to create a metering algorithm that would provide for indication of perceived loudness*", and "*that the state of digital signal processing makes it practical to implement complex algorithms into cost-effective devices*". Thus the Question 2/6 was posed, part 2 of which is to study: "*What audio metering characteristics should be used to provide accurate indication of subjective programme loudness?*", with the goal of basing a new ITU-R Recommendation on the results.

In order to decide on this question, a set of subjective reference data was produced by means of a loudness assessment experiment. In the experiments, a segment consisting of female speech was used as a fixed reference sound against which the other sound segments were matched in loudness. A total of 49 different sound segments were collected to be representative of broadcast programme material. In the collection, 13 of the segments are music, and the rest are primarily speech material with or without various background sounds. Listening experiments were conducted at 5 different test sites. Some details of the experiments and their results have been presented in [40, 41, 42].

In April 2003, a reflector for the SRG-3 was created at the Yahoo Groups web site, in order to allow all interested parties to participate[1] in the ongoing work [43]. (A number of the arguments presented in this paper (section 5) have already been put forward as contributions in the SRG-3 forum.)

Later in April 2003, a "*Call for submission of audio loudness metering methods*" was issued [44]. The response was a total of 10 loudness metering methods submitted by 7 different research organizations and private companies. The loudness algorithms implemented in the different meters were not disclosed to the SRG3. Some of the submitted loudness meters were implemented in hardware, others in software; some measured the loudness level in phon, others in relative dB. The official requirement of the meters was simply that they should provide a method of measuring the long-term loudness of a given sound segment, relative to a reference sound segment.

The proposed loudness meters were collected by the CRC [45], where they were then evaluated. Each model was used to measure the loudness of the same sound segments used as stimuli in the listening experiments. The criteria and methods of evaluation were discussed on the SRG-3 reflector, and the results of the evaluation were presented at the Montreal Meeting of ITU-R WP6P SRG3 [46], and was subsequently published in [42]. Two of the four models submitted by Dolby Labs were the L_{eq}(A) and L_{eq}(B), and during the meeting a technical description of the two others was provided. Briefly, the results of the evaluation indicated that two simple RMS- or energy-based loudness measures were superior to any of the submitted loudness meters, in estimating the relative loudness level of the 48 different sound segments. It has been discussed, within the SRG-3, whether the collection of sound segments, used in the experiments and the meter evaluation, was too narrow in terms of spectral and dynamic variation. The question

[1] Consequently, because non–ITU-members contributed to the work, the Special Rapporteur Group (SRG-3) was officially changed into a single Special Rapporteur, Mr. Craig Todd.

of whether certain bias factors might have been influencing the results, has also been raised.

At the time of writing, mainly part 2 of the Question has been addressed within the SRG-3 – namely, how to accurately measure the long-term loudness – and the results are inconclusive. A new round of listening experiments is being planned [47], to produce subjective reference data to complement (and validate) the reference data from the first round.

2 PREVIOUS LOUDNESS MODELS

Any loudness model can be categorized as either a single-band model or a multi-band model. The multi-band models split the input signal into multiple frequency bands that are subsequently combined into a loudness estimate, whereas the single-band models only have one signal path through the model. Ten of the loudness models, evaluated in this paper, are implementations of common or standardized objective loudness measures. The investigation included 8 different single-band models, based on two different principles: equivalent sound level (L_{eq}) and PPM level measurement. Two variants of the Zwicker loudness model were included as instances of a prevalent multi-band model. Sections 2.2-2.4 provide a brief structural description of each of these evaluated loudness models.

2.1 Types of Loudness Models

A classic question, related to modeling, is whether the model should reflect the inner structure of the (auditory) system, or should simply model the transfer characteristics between input and output variables (or terminals). Knowledge of a complex system can often be used to break down the system into simpler sub-systems, which may then be modeled using more manageable mathematics. Complex auditory models may cover the complete physical system from outer ear to the hair-cells [48, 49], and sometimes even cognitive effects.

If, in the case of loudness models, only certain types of input are considered, the model can be simplified dramatically. For example, the input might be limited to be just isolated pure tones. In this case, determining the loudness function by means of listening experiments, and subsequently computing the model parameters, is relatively straightforward [27]. In practice, a switching mechanisms might be used to adjust certain model properties, depending on certain properties of the input signal. An examples of this principle is a model that handles noise differently than tonal components [31, 50]. An application-specific model might be suitable just for speech signals and not for broadband music – or pure tones, for that matter. Such a model would be simpler than a more general one, but would often fail badly when used outside its intended scope.

Some earlier loudness models were designed to operate only on *stationary* signals, that is, sounds which can be described completely by their frequency spectrum, such as noises and tone complexes (e.g. [51]). Newer models attempt to also predict the perceived loudness of fluctuating signals [34, 35].

Although it is known that the inner ear performs a frequency analysis of the incoming sound, and that signal frequencies interact in a complex way [27], several *de facto* loudness models have been constructed without a frequency analysis (i.e. single-band models) [52, 53, 54, 55]. Some issues are common to both single-band and multi-band types of loudness models. For instance, in which domain – e.g. linear, squared, log – is time integration be handled.

2.1.1 Single-band methods

Several single-band models have been constructed over the years, based on the frequency-dependency of loudness perception and employing an appropriate envelope detector. A *frequency weighting* is typically based on an approximation of one of the equal loudness contours. As these contours vary as a function of level, a compromise must be made in the choice of weighting, e.g. the A- or B-weighting curves [56]. Spectral loudness summation (section 1.1.2) cannot be modeled using a simple broadband method. The typical use of single-band models is therefore in conjunction with broadband signals with similar spectral content, and within a relatively narrow level range.

The temporal properties of hearing can be modeled when using a suitable detector to account for the smaller loudness of short signals [34]. Sometimes, the long-term energy-integration measure L_{eq} is used in single-band methods. L_{eq} is a measure of the equivalent power, or RMS value, over time – from seconds to hours depending on the application. The duration dependent loudness (section 1.1.3), for tone and noise bursts are handled reasonably well by this energy-integration, even in single-band models. Especially when using a long-term average method, such as the L_{eq}, the

measurement can benefit from the use of a silence detector. Usually, a loudness measurement method is expected to estimate the loudness only when a signal is present, e.g. [53].

2.1.2 Multi-band methods

The discovery and quantification of the spectral loudness summation [33] has led to loudness models incorporating several frequency bands; some of them using critical bandwidth filters, some with fewer, and hence broader, filters in order to reduce complexity. Two pioneers in this area were S. S. Stevens and E. Zwicker. Although the multi-band models are generally more complex than single-band models, due to the extra task of filtering the signal into multiple bands, the multi-band approach provides some fundamental advantages in the accurate modeling of perceived loudness.

The method by Stevens is described in the standard ISO 532-A as being suitable for broadband signals without strong spectral peaks of large frequency separation [4]. Furthermore, a diffuse sound field is assumed and the sound should be steady-state. This latter condition is less limiting than it might seem, as a diffuse sound field will smear an impulsive signal in the time domain. The frequency analysis is basically performed in whole octaves, although the method can adapted to half- and third-octaves. The model comprises spectral loudness summation.

The method of Zwicker is also described in the ISO 532 standard, as method B [4]. In the standard a third-octave frequency analysis is assumed. This is a fairly good approximation to the critical bands. In the original work a filter bank with true critical bandwidth filters was used [51], but for technical and practical reasons, a third-octave filterbank was preferred. Compared to the method by Stevens, the Zwicker method comprises a more sophisticated spectral loudness summation including a spreading function closely related to the effect of simultaneous masking. Furthermore, frequency weightings for both diffuse and free sound fields are included. The method was indicated as suitable for both narrow and broadband signals.

Several researchers have contributed with refinements to the original Zwicker method [34]. In Cambridge (UK), Moore, Glasberg and colleagues have proposed another interpretation of critical bandwidth, and made several improvements, for instance for calculating the loudness of fluctuating sounds [35, 5, 50, 3].

Starting from the application side, work at the laboratories of the broadcaster CBS resulted in a loudness model based on octave-wide filters. It was constructed for measuring the loudness of signals typically encountered in radio and TV broadcasts [1]. A preliminary study [57] justify the properties of the model. The main starting point was Stevens' work, and also spectral loudness summation and time constants were included in the model.

A recent study [25] describes the effect on loudness caused by applying varying degrees of multi-band dynamics compression to speech signals. One of the interesting conclusions was that the long-term loudness, for a given signal, increases with the degree of dynamics compression applied, and with the RMS value held constant. It is generally expected that the RMS value increases for a given peak level when dynamics compression is applied. Loudness is often thought to be proportional to the RMS value (e.g. [55]), but the study [25] has demonstrated that even when the RMS value is kept constant the loudness depends on the degree of dynamics compression – or rather the resulting fluctuation depth. The loudness model described in [25] correctly models this effect of compression – at least to some extent.

2.1.3 Annoyance vs. loudness

It may be difficult to distinguish clearly between *loudness* and *annoyance*, as loudness is a major – but not the only – contributor to annoyance. Legislation regarding environmental noise is often based on the long-term A-weighted measurement, the L_{eq}(A), but with some corrections due to the known discrepancy between the basic L_{eq}(A) and perceived loudness or annoyance. Specifically, signals containing strong tonal or impulsive components are generally considered to be more annoying than more random signals of the same level. A Danish guideline for environmental noise [58] therefore indicates that 5 dB(A) should be added in the case of clearly audible tonal components or clearly audible impulsive components in the noise. The assessment of whether the sound is tonal or impulsive is specified to be subjective, due to the lack of an objective method. Similar principles apply in a recent EU directive [59] on environmental noise, but in which the exact amount of level to add, in case of tonal or impulsive character, is not indicated (work in progress).

2.2 Leq or RMS measures

The L_{eq} measure is the *equivalent continuous sound level*, or time-average sound level. The L_{eq} corresponds to an (energy domain) average over a time interval T during which the sound level is measured, in dB [60]. A mathematical definition of *Leq* is:

$$L_{eq}(W) = 10\log_{10}\left(\frac{1}{T}\int_0^T \frac{x_W(t)^2}{x_{Ref}(t)^2}\,dt\right)\text{dB} = 20\log_{10}\sqrt{\frac{1}{T}\int_0^T\left(\frac{x_W(t)}{x_{Ref}(t)}\right)^2 dt}\ \text{dB} \quad (3)$$

The $x_W(t)$ is the frequency-weighted sound pressure of the measured signal at time t, and $x_{Ref}(t)$ is the reference signal. Typically a frequency weighting W is applied to the measured signal, prior to the integration or averaging. The term *linear* L_{eq}, or $L_{eq}(Lin)$, refers to the unweighted L_{eq} as opposed to a frequency-weighted L_{eq}. The second formulation in eq. 3 is included to show the L_{eq} as a root-mean-square (RMS) type of measurement which is transformed into dB.

The L_{eq} is commonly used in acoustical measurements of sound sources with a time-varying level. Used together with certain frequency weightings the L_{eq} measure is often considered a measure of loudness. When the L_{eq} measure is employed as a model of long-term loudness, in this study, we simply assume that the measurement period T is equal to the duration to the sound segment to be measured. Note that although the L_{eq} was constructed as a *dose* measurement, it is only applied here for homogenous sound segments. The algorithm below shows the expression of the L_{eq} as one of the loudness models that we evaluated.

Algorithm for the *Leq(Lin)* loudness model:

1. calculate the unweighted L_{eq} measurement, in dB, of the entire sound segment (eq. 3)
2. apply calibration gain, so that a 1 kHz full-scale test tone corresponds to a fixed loudness level reference (such as 100 phon)

2.2.1 Leq (A, B, C, D, M, RLB)

The sensitivity of the human hearing is frequency dependent (e.g. [61]). Therefore a *frequency weighting* of the signal is commonly applied, prior to the calculation of the L_{eq}. Generally, such frequency weightings attenuate the low-frequency part of the spectrum, corresponding to the region where the hearing is least sensitive, roughly, below 100 Hz. Some of the weightings attenuate the high-frequency region as well, and may incorporate a peak in the frequency response around 1-4 kHz, where the hearing is most sensitive. As the sensitivity of hearing also depends on the absolute level of a sound, a given frequency weighting corresponds to a certain SPL range. Furthermore, the frequency weightings may serve different purposes: for instance, it may be applied to achieve an estimate of the perceived *loudness* or of the *annoyance* of a certain category of sounds.

In this study, 5 standardized frequency weightings have been implemented. The A-weighting is most commonly used with L_{eq} measurements [60]. The B- and C-weightings were constructed to complement the A-weighting, such that A-weighting should be used at low sound levels, B at medium, and C at higher sound levels [56]. The A-weighting was originally a simple approximation to a 40 phon equal loudness curve, and its although its correlation with loudness (and with annoyance) has been questioned [62], sound level measurements often report the dB(A).

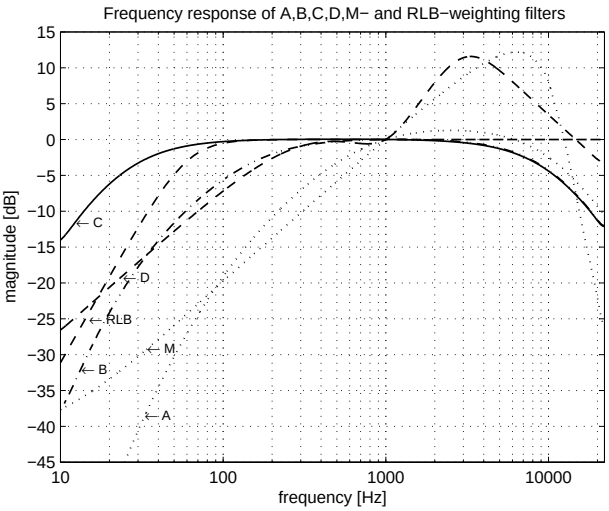

Figure 1. Frequency response of the A, B, C, D, M, and RLB frequency weighting filters. All the curves are aligned in level so that their magnitude response at 1 kHz is 0 dB.

The CCIR specified a weighting curve for the purpose of measuring low-level noise in electroacoustic devices, with a quasi-peak detector [63, 54]. However, this frequency weighting is also being used together with L_{eq} measurements. This so-called $L_{eq}(M)$ – the "M" is for

movie – has been promoted by Dolby, to be used for measuring the loudness of different segments of movie soundtracks such as advertisements [64, 52]. The frequency response of the A, B, C, D, and M-weightings are shown in Figure 1.

Soulodre and Norcross [55] evaluated different loudness models against subjective reference data. All the models were based on frequency weighted L_{eq}, and a new frequency-weighting was introduced: the Revised Low-frequency B-weighting (*RLB*). In the evaluation against the subjective data, the L_{eq}(RLB) was ranked higher than any of the other evaluated models. Note that the RLB frequency weighting corresponds to a *high-pass* filter, as opposed to the A, B, C, D, and M-weightings. A similar high-pass weighting has reportedly been used for some years by the Danish broadcaster *TV 2*, in connection with loudness metering [42]. The frequency response of our RLB filter implementation is shown in Figure 1.

Algorithm for the frequency-weighted *Leq* models:

1. apply the frequency weighting filter, *W*, to the sound segment
2. calculate the L_{eq} measurement, in dB, of the entire sound segment (eq. 3)
3. add calibration gain, so that a 1 kHz full-scale test tone corresponds to a fixed loudness level reference (such as 100 phon)

2.3 PPM + percentile

The *Peak Program Meter* (PPM) was developed as an alternative to the traditional VU meter. The PPM displays the pseudo-peak level of audio signals, using a fast attack time constant, allowing the meter to detect peaks, and a slow release time constant, to give the user time to read the meter's peak indication [6]. There are several PPM specifications with minor variations of the attack and release time constants. The implementation used in this evaluation was based on the German DIN 45406 standard (which is similar to IEC 268-18). The PPM is commonly used in the production of digital audio, where the engineer needs to monitor that the signal does not exceed 0 dBFS.

2.3.1 From envelope to long-term loudness

Unlike the L_{eq} measurement, which calculates a single average value, the PPM measurement returns a time-domain signal – an *envelope* in dB. Suppose an audio engineer was assessing the loudness of an audio signal, by watching the PPM, without hearing the sound. In this case the engineer and the PPM would together constitute a loudness model. For the purpose of our evaluation, we have replaced the engineer by a simple statistical function, to calculate an estimate of the long-term loudness given the envelope from the PPM measurement.

When the temporal dimension of the envelope is discarded, the data could be described by the *distribution* of dB values. One way of estimating the long-term loudness would then be to calculate a certain *percentile* of the distribution; for example, the 95th percentile is that value which the envelope is below 95% of the time. Spikofski and Klar considered the cumulative frequency distribution, based on histograms of PPM levels [65, 9, 10]. This approach is equivalent to the percentile method used here, because the cumulative distribution can be obtained by integrating over the envelope level distribution.

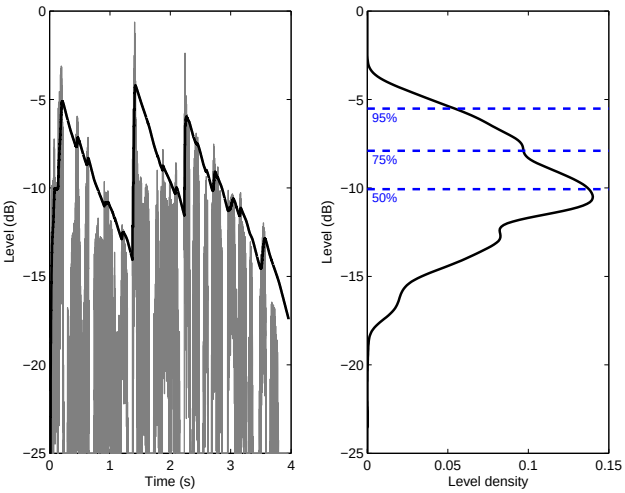

Figure 2. *Left*: The first 4 seconds of female speech; sample magnitudes (grey), PPM envelope (black). *Right*: The distribution or density of the PPM envelope levels; the 50th, 75th, and 95th percentiles of the distribution are marked.

The calculation of the long-term loudness estimate, based on the PPM envelope, was evaluated for three different percentile values (section 6.1): the 50th, 75th, and 95th percentiles of the distribution. To illustrate this procedure, Figure 2 shows the envelope produced by the PPM, for the first 4 seconds of track 53 (female speech) from the SQAM CD [66]. The figure shows the distribution of dB levels as a density function – in other words, the density plot on the right contains the same information as the envelope plot on the left, except that the time dimension of the envelope is collapsed. The

three different percentiles of the distribution are indicated in the figure.

In the algorithm below, steps 1-3 and 5 ordinarily belong to the PPM measurement.

Algorithm for the PPM+percentile loudness model[2]:

1. full-wave rectification of the input signal
2. asymmetrical low-pass filter, attack and release time constants as in the PPM
3. convert the output to an envelope in dB
4. calculate the p^{th} percentile of the distribution of envelope levels
5. add calibration gain

2.4 Zwicker Loudness models

In 1960, Zwicker presented a procedure for calculating the loudness of stationary sounds [51]. This highly influential model of loudness was based on a frequency division into critical bands [67] – hence the Zwicker model is a multi-band loudness model. The input to the model is the 1/3-octave spectrum of the sound, and the output is the estimated loudness in sone. The Zwicker model was adopted in the standards DIN45631 and ISO 532-B [4].

Appell *et al.* reviewed and measured the properties of four different multi-band loudness models [30]. It is described how these models, which are more or less based on the Zwicker model, have fundamental similarities in their structure and processing principles. The algorithm below outlines the processing steps of such Zwicker-type models.

General algorithm for Zwicker-type loudness models:

1. linear filter, to simulate the outer and middle ear frequency response
2. auditory filterbank, to simulate spectral masking (in Bark or ERB)
3. calculate the excitation pattern (excitation level in each filterbank channel)
4. calculate the specific loudness, using a power law compression, and accounting for hearing threshold
5. calculate the total loudness, in sone, by integrating the specific loudness across critical bands
6. (if required) transform the sone to phon (or some other loudness scale)

The original Zwicker model has been extended, for instance to better cope with time-varying sounds by modeling the post-masking effect [68, 27, 5]. Even though the Zwicker model is the most computationally complex model tested here, real-time implementations are available, e.g. [69].

2.4.1 Two separate implementations

For the purpose of our evaluation, an estimate of the long-term loudness of a sound must be calculated, given the time-domain signal or envelope, in sone, calculated by the Zwicker model. We have used two separate Zwicker model implementations, partly to test two different methods of calculating the long-term loudness estimate, and partly to detect any implementation-specific functioning.

1) SI++ implementation of the Zwicker model

The commercial acoustics analysis software, SI++, provides several implementations of the Zwicker loudness model [70]. We used the *lautheitII* function, which is implemented with a filterbank rather than with FFT (which *lautheitI* is). *LautheitII* accepts an audio signal as input, and computes the Zwicker loudness according to the ISO 532-B standard, furthermore incorporating *post-masking* (which is not specified in the standard). The function was used in *free field* mode, with a specified output rate of 50 Hz.

Zwicker and Fastl presented the N_5 loudness which is the loudness, in sone, that only 5% of the loudness values are above, in the cumulative loudness distribution [27] (pp.318-324). The N_5 was used to predict the subjective overall loudness of noise emissions, based on a Zwicker loudness model. Moreover, the documentation accompanying a commercial implementation of the Zwicker loudness, states that, "*The perceived loudness of a long, non-stationary sound is the loudness value which is exceeded 5 % of the time in the loudness/time course.*" [71]. This statistic corresponds to 95^{th} percentile.

We have chosen to estimate the long-term loudness using a percentile statistic of the distribution of sone values, analogous to the procedure we used with the PPM measurement (section 2.3.1). We shall denote this model "Zwicker&Fastl(95%)", when the 95^{th} percentile is used. Like in section 2.3.1, three different variants the model were evaluated, with different percentile values.

[2] For efficiency reasons steps 3 and 4 could be interchanged; the percentile statistic is invariant to transformation by any monotone increasing function.

Algorithm for the *Zwicker&Fastl(percentile)* loudness model:

1. scale the sound segment, so that the signal represents the corresponding SPL
2. call the *lautheitII* function implemented in the SI++ system
3. transform the sone envelope to phon
4. calculate the p^{th} percentile of the distribution of loudness levels

The transformation from sone to phon used the relation in eq. 4, together with a floor function (which had no effect on the result due to the percentile statistic). In eq. 4, P is the loudness level in phon, and S_t is the total loudness in sone [4].

$$P = 40 + 10 \cdot \log_2(S_t) \tag{4}$$

2) ISO 532-B implementation of the Zwicker model

An alternative implementation of the Zwicker model was based on the C code by Widmann [72][3]. This function implements the procedure specified in the ISO 532-B standard. The C code was itself based on a QuickBasic program by Zwicker *et al.* [73], which again was based on a Fortran program by Paulus and Zwicker [74]. We shall denote this model "Zwicker ISO".

The ISO 532-B procedure calculates the loudness given a set of third-octave levels. In an acoustic measurement, third-octave band L_{eq} levels are typically obtained from a spectrum analyzer (e.g. [75]). We used this L_{eq} approach, even though the sounds in our evaluation are non-stationary. Thus, the time-dimension in the "Zwicker ISO" implementation is collapsed *before* the calculation of the loudness, whereas in the "Zwicker&Fastl(95%)" implementation, it was collapsed *after* the loudness calculation (via the percentile statistic).

Algorithm for the *Zwicker-ISO* loudness model:

1. apply calibration gain, so that the digital representation of a 1 kHz test tone corresponds to the appropriate loudness level in phon
2. third-octave filterbank
3. calculate the L_{eq} (in dB) per frequency band
4. perform the three-step procedure specified in ISO 532B, using the C implementation
5. transform the resulting total loudness in sone to phon

[3] Minor (seemingly typographical) errors in the C code were corrected, to make the program and data consistent with the BASIC version of the code.

2.5 Other models (not evaluated here)

The CBS Loudness Indicator [2] is a multi-band loudness model, based on a filterbank with 8 bands each covering 3 critical bands. This model has served as a *de facto* reference for objective loudness measurement, in the broadcast community.

Moore and Glasberg has developed a multi-band loudness model, structurally comparable to the Zwicker model, but with certain enhancements [50]. Recently, the model has been expanded to predict the long-term loudness of amplitude modulated (i.e. time-varying) sounds, via the computed short-term loudness [35]. With a minor adjustment, this model was also able to predict the long-term loudness of speech stimuli treated with various degrees of dynamic compression [25].

It would have been interesting to include, for example, the CBS Loudness Indicator and the revised Moore & Glasberg model in this evaluation. Unfortunately, to our knowledge no (software) implementations of these models were publicly available. Stevens method, as standardized in ISO 532-A, was also not considered.

3 TWO NEW MODELS OF LOUDNESS

Our initial evaluation of loudness models indicated that none of the available models were able to provide an accurate estimate of the perceived loudness of music and speech segments. Moreover, none of the existing loudness models were originally designed to estimate the loudness of non-stationary sounds such as music; although, in practice, they are frequently utilized for precisely this purpose.

At TC Electronic, we have therefore developed two new models of loudness. The main objective of the new models has been to produce an accurate and robust estimate of the perceived loudness of sound segments consisting of speech and/or music. In the following, we introduce a new single-band loudness model and a new multi-band model. Both these models were optimized to compute an accurate long-term loudness estimate. However, they can also be used to compute a short-term loudness, by shortening their analysis-window.

3.1 The new single-band model: LARM

The new single-band model, named *LARM*, consists of a combination of two known measurement techniques: the pseudo-peak detector, and the frequency-weighted RMS

measurement. An asymmetrical low-pass filter with a short attack time-constant and a longer release time-constant is used in the PPM to emphasize the higher levels in the envelope. This property is also relevant in a loudness model (see section 1.1). The LARM model calculates the loudness estimate analogous to the frequency-weighted L_{eq} (eq. 3). However, as a generalization of the RMS calculation, a variant of the *power mean*, M_p, is employed (eq. 5), such that p is in effect turned into a model parameter.

$$M_p(x) = \sqrt[p]{\frac{1}{N}\sum_{i=1}^{N}|x(i)|^p} \tag{5}$$

Algorithm for LARM:

1. apply frequency-weighting filter to the sound segment
2. full-wave rectification
3. asymmetrical low-pass filter (i.e., separate attack and release time constants)
4. calculate power mean, *Mp*, of the envelope levels for the entire sound segment
5. transform the *Mp* into dB
6. add calibration gain, so that a 1 kHz full-scale test tone corresponds to a fixed loudness level reference (such as 100 phon)

The LARM model contains several model parameters that may be tuned to achieve the most accurate loudness estimate. In doing so, we have made the following observations: Step 1 could be a simple high pass filter, like the RLB-weighting, but the frequency response may be further optimized. In step 3, the best performance is obtained when the *release* time-constant is considerably slower than the *attack* time-constant, i.e. an asymmetrical low-pass filter which accentuate peaks. In step 4, any M_p, for $p > 1$ represents a non-linear weighting of the envelope values resulting from step 3. Note that for $p = 2$, step 4 is equivalent to calculating the RMS value, and for $p = 1$, step 4 essentially calculates the mean absolute value. Curiously enough, we have for LARM found an optimal value of $p \approx 1.5$.

3.2 The new multi-band model: HEIMDAL

About 45 years ago Zwicker proposed a loudness model founded on the concept of critical bandwidth (section 2.4). Subsequently, most multi-band models of loudness have used third-octave or auditory filterbanks to simulate the spectral masking (section 2.1.2). For speech and music signals, however, there is a strong interdependency between the dynamical behavior at frequencies in neighboring critical bands. For these types of signals, it may therefore be possible to model the loudness, using a resolution less than the critical-band rate.

The new multi-band model, named *HEIMDAL*, comprises a novel algorithm based on an octave-band filterbank. The input signal is filtered into 9 channels which are then processed individually. In the HEIMDAL model, octave band filters are used as a compromise between computational complexity and spectral resolution. By using a filterbank and not an FFT analysis, the time/frequency trade-off associated with the latter is avoided. It is beyond the scope of this paper to describe the details of the HEIMDAL model, but a structural outline is provided below.[4]

Algorithm for HEIMDAL:

1. rms-based normalization of signal level
2. octave-band filterbank
3. full-wave rectification
4. asymmetrical low-pass filter
5. extraction of multiple features per frequency band
6. non-linear combination of features into estimate of relative loudness
7. calculate long-term loudness estimate
8. add calibration gain

3.2.1 Model profiles, optimization and generalization

The HEIMDAL loudness model contains a set of model parameters, corresponding to step 4, 5, and 6 in the algorithm. These model parameters have been optimized so that the model will produce the most accurate loudness estimate for a range of different sounds. The values of the model parameters are determined during the optimization phase, and will not need to be adjusted in the application of the model. On the other hand, if a different loudness function was desired, a new set model parameters – a new *profile* – could be installed into the HEIMDAL model.

Determining the values of the model parameters is essentially a non-linear optimization process. This process involves a trade-off between the accuracy of the model's predictions of the target values, and the model's ability to accurately estimate the loudness of 'new' input signals. The objective of the optimization is for the optimized model to produce accurate estimates for any new input of the same type as that used in the optimization. This property is known as the *generalization ability* of the model. The generalization

[4] International patent pending.

(error) can be tested and improved using techniques from machine-learning, e.g. [76, 77].

When using the same data to optimize the model parameters and to estimate of a model's generalization ability, the estimate will tend to be optimistically biased. Thus, it is preferable to evaluate the model's performance using data which is *independent* of that used for optimization. In case not enough data is available to constitute two independent, representative data sets, *cross validation* can be used: in an iterative process, the available data is partitioned into different subsets of data, used for optimization and testing, respectively.

In principle, three independent sets of data are required for the optimization procedure: The *training set* is used for the actual optimization of the model parameters. The *validation set* is used to ensure the generalization ability of the model, and to choose between different candidate models. The *test set* is used to finally measure the performance of the now fully-specified, optimized model [77, 76].

Two independent data sets were available for the model evaluation presented in this paper: One data set was used for the optimization of the LARM and HEIMDAL models, in combination with a cross-validation technique; the other data set was used as the validation set.

3.2.2 Limitations of the current HEIMDAL implementation

In this evaluation, we have chosen not to include the type of stimuli traditionally used to develop and test psychoacoustic loudness models, such as pure tones and noise bursts. The HEIMDAL model could to some extent predict the 'correct' loudness level of such sounds – this aspect might be addressed in a subsequent paper. However, due to the octave-band resolution, signals could be constructed, for which HEIMDAL's loudness prediction would be less accurate than that of the psychoacoustic loudness models. On the other hand, such signals are quite rare in typical programme material which generally has a fairly continuous spectrum.

The current implementation of HEIMDAL is neutral to the absolute SPL, in the sense that if the stimulus is turned down by x dB, then the loudness estimate will likewise drop by x phon. This behavior, caused by the normalization in the algorithm's step 1, contradicts the fact that the frequency sensitivity of the hearing is known to depend on the SPL of the stimulus. Furthermore, this HEIMDAL implementation will in principle have an infinite dynamic range, and will not simulate the threshold of hearing. In practical applications of the model, in connection with audio production and broadcast, this limitation may be less of a problem because the loudness level of most sounds tends to be in the range 55 to 95 phon. In some psychoacoustic models, the threshold is simulated partly by adding a constant low noise to the stimulus, thus masking the excitation produced by the stimulus below the threshold [30]. One extension to the HEIMDAL model could be to incorporate a similar technique, in effect providing an appropriate loudness 'floor'.

4 SUBJECTIVE LOUDNESS REFERENCE

Loudness is a perceptual and subjective property, and as such it is more difficult to measure than physical properties. Both uncertainty and subjectivity factors have to be taken into account. An evaluation of loudness models requires a data set of representatively sampled sound segments together with their corresponding perceived loudness levels. Currently, no such data set seems to be available to investigations such as ours. In this respect the development of loudness models is lacking behind for instance *perceptual codec* development, for which standard evaluation data and meticulous procedures are available to the developers [78]. In principle, a data set and evaluation procedures could similarly be compiled for the purpose of a consistent loudness model assessment, by a standardization organization such as ITU-R, AES, or EBU. The lack of 'standard' subjective reference data and methodology, combined with the difficulty and expense of obtaining reliable results via listening experiments, makes it difficult to compare the results of different loudness model studies.

Established loudness models have sometimes been assumed to provide subjective reference data against which *other* models could be evaluated. For instance, Benjamin explicitly states "*...the assumption that the Zwicker or Moore loudness measurement methods represent the real perceived loudness of the program samples*", when using these two models as reference in his study [79]. In [80] the CBS Loudness Indicator (meter) was used as reference because "*This instrument indicates true perceived loudness and agrees well with a panel of human listeners.*".

4.1 Loudness experiment method

The subjective reference data, used in this model evaluation, was derived from two listening experiments, using the speech and music segments as stimuli. The two experiments were conducted at different sites, with different test subjects and stimuli, resulting in two independent sets of reference data. The experimental method and statistical analysis, used in both experiments, were presented in [23], and is summarized in the following.

4.1.1 Balanced pair-matching

The *loudness matching* experiments were conducted using the *method of adjustment* [27, 81]. The subjects were instructed to adjust the loudness of a comparison stimulus (B) using a volume or gain control until it matched a reference stimulus (A). The relative level of one of the segments in each pair was controlled by the subject using an endless rotary knob [82]. The method of adjustment was chosen because it was fast and intuitive, and therefore suitable for an experiment involving a relatively large number of segment pairs to match.

Measuring the perceived loudness in a listening experiment implicates several choices regarding the experimental design and procedure. In particular, a loudness matching experiment could be based on a *fixed reference* method, in which the subjects match all stimuli against a single sound segment selected in advance. We have employed a different method, however: the ***balanced pair-matching*** method. When using this method, both segments in a pair are drawn from the same collection. No single fixed reference sound is used in the balanced pair-matching method, and thus the influence (bias) that a specific reference sound could have on the results is "evened out". The composition of the set of pairs to be matched is said to be *balanced*, because the relative frequency of occurrence of the different segments is the same [24].

The experimental design was constructed with redundancy, in the sense that each subject performed more than one loudness match involving each sound segment. By affording a certain redundancy and by using a balanced experimental design, the accuracy of the results for the balanced pair-matching method was improved beyond the fixed-reference method with any segment as reference [24]. The results of a pilot experiment indicated that the improved accuracy was caused partly by using the balanced pair-matching method, and partly by the limited redundancy in the design.

4.1.2 Stimuli

The sound stimuli used for the experiments consisted of monophonic segments of speech and/or music. The choice of monophonic stimuli was made to eliminate possible extra factors introduced by stereo signals that could have affected the loudness assessment of the subjects. The segments were extracted from commercial music recordings, radio broadcasts, and movie soundtracks. Each segment was edited into a short excerpt of approx. 10 to 15 seconds in duration. Each segment was selected to be homogeneous with respect to its spectral content, dynamic properties, and instrumentation to facilitate the assessment of its overall loudness. Each segment was normalized in level by means of an RMS-based estimate of its loudness. A random offset was the added to the presentation-level; the variance of these offsets thus determined the 'dynamic range' of the experiment.

4.1.3 Statistical analysis

In [23], a statistical model of the responses from a *balanced pair-matching* experiment was presented, in the form of a General Linear Model (GLM) [83, 84]. The statistical model incorporates bias terms corresponding to two types of errors associated with the method of adjustment: the adjustment bias and A/B-order bias. Both types of bias were found to be significant (in the modeled responses) for 3/4 of the subjects. By including bias terms for each subject, in the statistical model, the biases can be estimated and removed. Furthermore, the redundancy incorporated in the experimental design was used to reduce the influence of the subject's inconsistency on the results. In the statistical model, all adjustments are used for estimating the loudness level of each individual sound segment. Hence the resulting loudness estimates would minimize the overall error [24].

In the GLM, the ***SegmentLevel*** variable comprised the loudness level for every sound segment in the experiment. Given the responses from the listening experiment, the GLM would estimate the optimal and bias-corrected *SegmentLevel* parameters. The set of *SegmentLevel* parameters can be estimated using the combined responses from *all* the subjects taking part of the experiment – in this case the *SegmentLevel* values

were said to describe the ***common loudness***. In other words, the *common loudness* estimate of the *SegmentLevel* of a given segment is the single best estimate of that segment's loudness level, based on all of the adjustments obtained in the experiment. Alternatively, the adjustments from each subject individually could be used to estimate the set of loudness levels. In this case the estimates were said to describe the ***subjective loudness***, and the parameters are denoted $SegmentLevel_S$. Thus, the adjustments of a given subject have no influence on the $SegmentLevel_S$ estimates of any other subject. The *subjective loudness* estimates are required when we wish to evaluate the predictions of a loudness model, *relative* to the between-subject disagreement [23]. The GLM contained individual bias terms, both when estimating the *common loudness* and *subjective loudness* parameters.

In summary, the experimental method outlined above, was used to obtain estimates for the loudness level for each sound segment. A statistical analysis was used to compute estimates of the *subjective loudness*, based on the adjustments of each individual subject, and to compute estimates of the *common loudness*, based on all adjustments obtained from all subjects in the experiment. The two types of loudness estimates provide the subjective reference data against which we evaluate the different loudness models.

4.2 Two independent experiments

The subjective reference data used in our evaluation were derived from two different listening experiments: One of the experiments was conducted at the McGill University by Dr. René Quesnel, and is described in [23]. The other experiment was conducted at TC Electronic, using an identical experimental method, but with a different set of stimuli (Table 2). Also the test subjects, loudspeakers, and listening rooms were different in the two experiments (Table 1). The subjective data from the two experiments can thus be considered to be *independent*, which is desirable for the purpose of model evaluation.

In the experiments, the *reference level* denotes the sound pressure level which the stimuli were centered around (see the level normalization, section 4.1). We have chosen a reference level of 70 dB SPL, as a compromise between domestic (TV) sound level [85, 16, 86], and public presentations, for instance in the cinema (the SMPTE RP-200 standard [87]).

The *standard error* of the *SegmentLevel* estimates is considerably lower in the McGill experiment than in the TC experiment (Table 1). This difference could be caused by several reasons: 1) the McGill subjects were all expert listeners, 2) the subjects of the TC experiment only performed half as many adjustments, i.e., less data to reduce experimental error, 3) the TC experiment contained a number of test tones and noises among the stimuli, some of which were difficult to match in loudness, hence yielding more uncertain adjustments.

A subset of the sound segments used in the two experiments was included in the reference data sets (Table 2). These segments were selected from four broad but distinctive classes of sound: speech, classical music, rock/pop, and jazz. The speech class included isolated speech and dialog, and speech with background music and/or environmental sounds. The music classes

	McGill University loudness experiment	**TC Electronic loudness experiment**
Number of test subjects	8	8
The subjects' background	Expert listeners, experienced in level adjustment	Ranging from 'naive' to expert listeners, with varied music-related backgrounds
Total number of adjustments submitted per test subject	1073	504
Total effective duration of experiment	37.8 subject hours	21.5 subject hours
Total number of sound segments used as stimuli in the experiment[1]	147	129
Type/source of audio segments	Ranging from raw mixes or un-processed recordings to final masters	Mastered and broadcast material
Standard error of the *common loudness SegmentLevel*	0.18 dB	0.43 dB
Reference level (around which the stimuli are centred)	70 dB(C) SPL	70 dB(C) SPL
Loudspeakers	Genelec 1031A	Dynaudio BM-15A
Loudspeaker setup	Stereo setup	Stereo setup
Distance from listener to loudspeakers	1.6 m	1.8 m
Listening room size and character	Small, damped	Medium, damped

Table 1. Parameters of the two listening experiments, from which the subjective reference data is obtained.

included both instrumental and vocal music segments.

A number of segments were excluded from this evaluation, because they were synthetic sounds, such as pure tones or filtered noise. The scope of this paper is the type of material that is likely to occur, for instance, in radio/TV programmes. Additionally, certain segments were excluded from the TC experiment because they also appeared as stimuli in the McGill experiment, and would thus introduce an overlap.

Sound segments in reference data set:	McGill University loudness experiment	TC Electronic loudness experiment
Rock or pop music	40	11
Jazz music	40	0
Classical music	40	0
Speech material	25	37
Total segments used in the reference data set	145	48

Table 2. The classes of stimuli from the two listening experiments, that are used in this model evaluation.

4.3 Principal component visualization of the reference data sets

Principal component analysis (PCA) is a common method for visualizing multivariate data [88, 89, 90]. We have used PCA to investigate the distribution of the sound segments forming the two reference data sets. Via the PCA, a *map* of the segments is produced, so that we could inspect whether the data sets constitute representative samples, and whether the two data sets overlap or have 'holes'.

A PCA was computed on the loudness features extracted by the HEIMDAL model, for the sound segments in the two collections (Table 2). Specifically, the PCA used the input features to the nonlinear function computing the loudness, i.e., extracted from between step 5 and 6 of the HEIMDAL algorithm (section 3.2). The covariance of the normalized features was used in the PCA. The HEIMDAL model utilizes the information provided in these features to estimate the relative loudness level of a given sound segment. The HEIMDAL performs an L_{eq}-type level normalization of the input signal. Therefore, the HEIMDAL features, used in the PCA, can be said to characterize aspects of a sound which are related to its loudness, but which are not captured by the L_{eq} measurement.

The first two principal components alone account for roughly 60% of the variance in HEIMDAL features for the combined data sets. In other words, 40% of the variance in the HEIMDAL features is lost by projecting them onto a two-dimensional plane. By considering only the first two principal components, however, the data is of course much easier to plot. The two most principal components of each sound segment in the McGill and the TC reference data sets are plotted in Figure 3. Additionally, the pink noise segment and the 1 kHz tone are included as fix points in the plots (these segments were *not* part of the reference data sets used for the model evaluation).

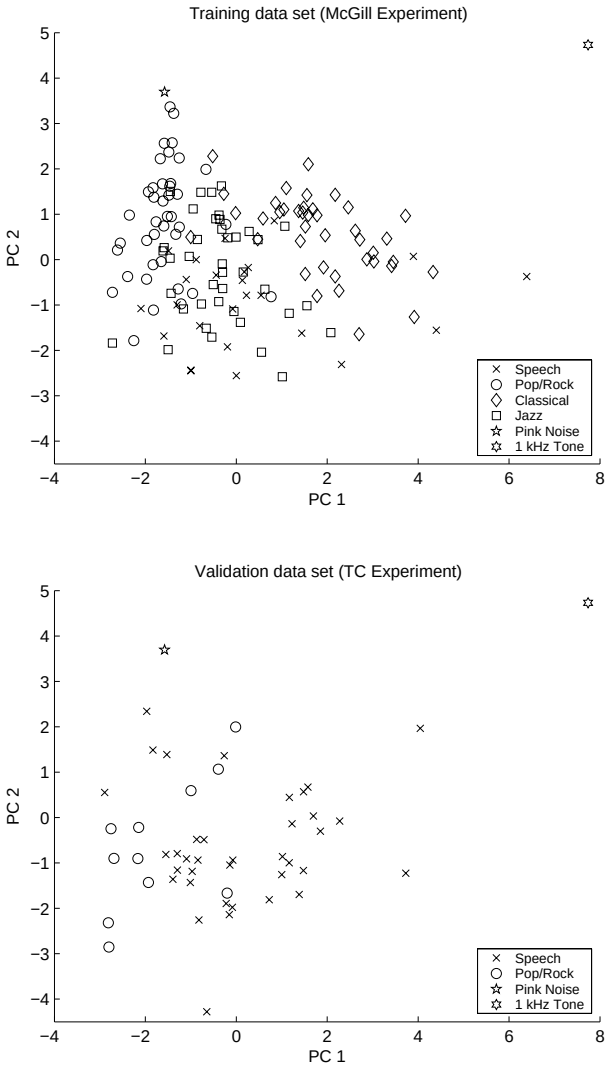

Figure 3. The two most principal components of each sound segment in the McGill (*top*) and TC (*bottom*) reference data sets. In each plot the pink noise segment and the 1 kHz tone are added as fix points.

By considering the sound segments in terms of the two most principal components of the HEIMDAL features, we can get an idea of how different the segments in the two collections are, as seen from a HEIMDAL perspective. In the context of a loudness model evaluation, we might use this kind of analysis to indicate how general or representative the results of the evaluation are. Even though the PCA projection 'spreads out' the data in the 2D plane, any large 'holes' in the projection of the segment collections could mean that a certain type of sound was not represented.

Although the different number of segments in the two collections makes it difficult to judge, it appears that the TC collection is 'narrower' than the McGill collection (Figure 3). It seems that the McGill collection extends longer than the TC collection in the positive direction of both PC-1 and PC-2 (the two principal components). What the TC collection lacks seems to be largely the classical music segments, a lot of which are centered around the coordinates (3,0). Furthermore, the McGill collection seems to contain pop/rock segments which reach all the way 'up' to the pink noise. This might be caused by the inclusion of unmastered material in the McGill collection. On the other hand, the TC collection contains several speech segments in areas where the McGill collection has none.

In summary, the two segment collections do overlap, in the plane spanned by the two most principal components of the HEIMDAL features. The overlap is not complete, however, because the McGill collection consists mainly of music, both mastered and unprocessed, while the majority of the TC collection is processed speech material, such as commercials or film sound tracks. The 1 kHz pure tone was isolated and not adjacent to the music/speech material of either segment collection. This separation highlights the difference between 'ecological' sound and test tones – also from a loudness model's perspective.

5 MODEL EVALUATION

Different statistical measures were employed in the evaluation of the models, so that both the absolute accuracy of the models and the accuracy relative to the subjective variability are measured.

5.1 Subjective loudness or Common loudness as Target

A loudness assessment experiment produces for each sound segment a loudness level corresponding to each test subject. There are two different approaches to evaluating a loudness model against this set of loudness levels:

1) **using the *common loudness* as target**. For each sound segment, a *single* target level is calculated. This *common loudness* value could simply be the average or median of the individual loudness levels, submitted by the different subjects. Alternatively, a statistical model could be used to estimate the *SegmentLevel* parameters which best fit all the adjustments obtained from all subjects (see section 4.1).
2) **using the *subjective loudness* as target**. In [23], we found a between-listener disagreement that was greater than what could be explained by the within-listener inconsistency. In other words, even for an ideal (infinitely large) listening experiment, with zero experimental error and an infinitesimal within-subject inconsistency, we could still expect a between-subject variability in the assessed loudness levels for each stimulus. In this case, the subjectively "correct" loudness level for a given sound would be a range or *distribution* of loudness levels, instead of a single level. The between-subject disagreement could be caused by individual factors, such as hearing ability and musical taste, or external factors, such as the acoustics of the listening room and loudspeakers used in the experiment. It might be appropriate to eliminate some of these subject-specific bias factors from the *subjective loudness* levels, prior to using them as reference for the models.

Note that performing a larger experiment (more adjustments) does *not* in principle lead to a smaller spread in the distribution of the *subjective loudness* for each sound segment, when that spread depends on between-listener disagreement rather than within-listener inconsistency. A larger experiment should lead to a better estimate of the distribution of *subjective loudness*, but the distribution's spread or range would remain the same. This is in contrast to the estimation of the *common loudness* – the more ratings or adjustments obtained in the experiment, the more accurate the *common loudness* estimate would be. Therefore, when comparing their results, researchers should be careful

not to confuse the spread in *subjective loudness* with the uncertainty of the *common loudness*.

5.2 Zero-order correction

Suppose the loudness levels predicted by a certain model would consistently differ from the target *SegmentLevel* by a constant value. This situation might be caused by the model not having been calibrated to the subjective reference data. This type of calibration problem could be corrected by simply adding an offset to the model predictions, so as to remove the systematic error. The loudness meter evaluation, within the ITU-R SRG3, applied this simple mapping to all meters – it was called the *zero order correction* [46]. A zero-order correction has also been applied to all loudness models evaluated in this paper, prior to measuring their accuracy. Note that the zero-order correction is specific to each set of subjective reference data.

If the $ModelPrediction_{uncal}$ is the prediction from an *uncalibrated* loudness model, the zero-order correction could be applied by subtracting the difference between the mean model prediction and the mean *SegmentLevel*.

$$ModelPrediction(i) = ModelPrediction_{uncal}(i) - \frac{1}{N}\sum_{i=1}^{N}\left(ModelPrediction_{uncal}(i) - SegmentLevel(i)\right) \quad (6)$$

5.3 Measures and statistics of evaluation

Numerous methods could be devised to evaluate the loudness models, either using visual displays or numerical measures of their accuracy. A numerical measure could quantify the error of the model predictions, compared to the *common loudness*, or relative to the distribution of *subjective loudness*. In addition, some measures use meaningful units, and furthermore, some measures can be compared across different experiments.

5.3.1 Measures of fit and residual error

Traditionally, a model's goodness-of-fit is measured by a summary of the residual error, and also the correlation coefficient (e.g. [83]). The *SegmentError* of a model prediction, compared to the *common loudness* estimate, for the i^{th} sound segment is -

$$SegmentError(i) = ModelPrediction(i) - SegmentLevel(i) \quad (7)$$

To summarize the error, across the entire segment collection, the *mean absolute error* and the *root mean squared error* are commonly used.

$$AAE = \frac{1}{N}\sum_{i=1}^{N}\left|SegmentError(i)\right| \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}SegmentError(i)^2} \quad (9)$$

The AAE[5] and RMSE measures both describe the overall absolute error, in dB. The difference is that the RMSE emphasizes the influence of relatively large errors over smaller errors, whereas the AAE is a simple average. Neither measure reflects the *uncertainty* of the *SegmentLevel* values, and thus inaccurate *common loudness* estimates may cause the AAE and RMSE to be systematically too large, i.e. biased.

Soulodre proposed to also measure the *maximum absolute error* [55]. Because this measure essentially ignores all *SegmentError*s but the single largest, it will tend to fluctuate a good deal for different sound collections. As a more *robust* measure of assessing the worst case predictions of a model, we propose the *95th percentile absolute error* (*P95AE*). Moreover, we found a visual display of the absolute error *distribution* to be very informative – in particular the *upper tail* of the distribution represents the *maximum absolute error*.

$$P95AE = \text{the value which 95\% of the absolute } SegmentError \text{ is below, and 5\% is above} \quad (10)$$

Pearson's correlation coefficient measures the strength of linear association between two variables. Suppose we have a new *SegmentLevel* and *ModelPrediction* variable, but with *relative* levels, such that each new variable has a zero mean value:

$$SegmentLevel_Z(i) = SegmentLevel(i) - \overline{SegmentLevel} \quad (11)$$

$$ModelPrediction_Z(i) = ModelPrediction(i) - \overline{ModelPrediction} \quad (12)$$

[5] In order to avoid confusion, we have adopted Soulodre's naming-convention [55, 42], and use the term Average Absolute Error, *AAE*, rather than Mean Absolute Error, *MAE*. (MAE was used by Soulodre to mean *Maximum* Absolute Error).

Then the correlation coefficient, R, is calculated as -

$$R = \frac{\sum SegmentLevel_Z \cdot ModelPrediction_Z}{\sqrt{\sum SegmentLevel_Z{}^2 \cdot \sum ModelPrediction_Z{}^2}} \qquad (13)$$

where the indices have been omitted, for clarity.

The correlation coefficient is invariant to any monotone increasing linear transformation of the data. Thus the correlation might evaluate a loudness model positively, even though it predicted the *common loudness* in units that were somehow scaled. In other words, the correlation coefficient implicitly performs a *first-order correction* in addition to the zero-order correction (section 5.2).

The *specific* correlation coefficients, obtained from a given model evaluation, are partly determined by the spread or variance in the reference data. Imagine the extreme case, in which all the sound segments were loudness-equalized by a (hypothetical) perfect loudness function, prior to the evaluation. In this case, the more a loudness model would deviate from predicting the (correct) constant loudness level, the more *negative* the correlation would be. Now imagine the other extreme, where the sounds were attenuated or amplified such that their loudness levels would span a range of say 100 dB. In this case, even the worst of loudness models would probably achieve a correlation coefficient close to 1.0.

The covariance between $SegmentLevel_Z$ and $ModelPrediction_Z$, calculated in the numerator in eq. 13, can be dominated by values which are unusually large, positive or negative – the so-called *outliers*. Just one outlier can be totally responsible for a high correlation coefficient that would otherwise have been close to zero [83]. A rank-order correlation (*Spearman's rho*) was used in [46, 42] to 'validate' the correlation coefficient, R. In this study, the confidence intervals are computed for the correlation coefficient, which would reveal the influence of outliers.

Another approach is to base the evaluation measure on a specific type of *application* of the loudness model. For instance, in case the model was intended to automatically balance the loudness of programme material, prediction errors less than 1.25 dB might go unnoticed, whereas errors larger than 10 dB may be unacceptable [55, 42]. With this in mind, Soulodre proposed the Loudness Performance Index (*LPI*) [55].

$$LPI = \prod_{i=1}^{N} \max\left\{1 - \left(\frac{|SegmentError(i)|}{L}\right)^p, 0\right\} \qquad (14)$$

Like the AAE and RMSE, the LPI calculates a single score based on the set of *SegmentError*, but differs in two respects: 1) the LPI weighs the error according to two constants, L and p, and 2) the resulting scores are multiplied (not added). The recommended value of L is 10 dB, which means that any model with one or more *SegmentError* greater than 10 dB would yield a total LPI of zero – the minimum score. With the suggested value of $p = 2.5$, any *SegmentError* of less than around 1 dB would receive only a small penalty.

As the outcome of the LPI depends entirely on the two design constants, the values of which are somewhat arbitrary, this measure was not used in our evaluation.

5.3.2 Two measures of Subjective Deviation

The error of a model prediction may be assessed, *relative* to the *subjective loudness*. This way, the *between-listener disagreement* is taking into account.

One statistical approach is to consider whether the model predictions are located within the *95% confidence interval* around the mean of the *subjective loudness*, for each sound segment. All predictions inside this interval are considered equally good, and only predictions outside the interval are counted as errors. The percentage of errors could then be used as an evaluation measure. This method can be regarded as a set of statistical hypothesis tests of the form: does the test reject the null hypothesis that the mean *subjective loudness* and the model prediction is the same value? If the null hypothesis is rejected, at the given significance level, for a sound segment, a significant error is counted. This method corresponds to the *tolerance interval* suggested as a minimum model requirement by Opticom [91], in relation to the SRG-3 work.

The confidence interval approach yields only a binary evaluation for each sound segment, and thus implies a qualitative difference between a model prediction just inside, and one just outside this interval. To obtain a more gradual evaluation, we proposed a new evaluation metric: the *Subjective Deviation*.[6]

[6] We first presented the Subjective Deviation formulae in the context of the ITU-R SRG-3 work [46].

For each sound segment, *i*, in the test sound collection, the Subjective Deviation is defined as -

$$SubjDev(i) = \frac{\left|ModelPrediction(i) - SegmentLevel(i)\right|}{InterQuartileRange\left(SegmentLevel_S(i)\right)} \quad (15)$$

The denominator in eq. 15 is the inter-quartile range of the *subjective loudness*, for the sound segment whose level is estimated. Thus, the Subjective Deviation measures the error of the model prediction, in a unit which is the spread of the 'middle' half of the test subjects. As the most disagreeing upper and lower 25% of the subjects are excluded in the estimation of the subjective spread, the estimate is *robust* against various aberrations. The robustness of the inter-quartile range is the reason it was preferred over the standard deviation which could alternatively have been used.

If the loudness of a certain stimulus was difficult for the subjects to assess, the models are punished less hard for deviating from the corresponding *common loudness* level. Note that the Subjective Deviation implicitly characterizes both the performance of the evaluated model and also the quality of the subjective reference data. The wider the spread in the *subjective loudness*, the more tolerant the Subjective Deviation measure would be.

The distribution of the Subjective Deviation may be characterized by means of a histogram or a box plot. Alternatively, two measures are presented, to summarize the Subjective Deviation as a single score. The *SDmean* is simply the average Subjective Deviation over the reference data set.

$$SD_{mean} = \frac{1}{N}\sum_{i=1}^{N} SubjDev(i) \quad (16)$$

The *SDprod* is a penalty-based measure, inspired by the LPI. The *SDprod* is a normalized product of factors between 0 and 1. Each sound segment will yield a factor close to 1, if the model prediction had a low Subjective Deviation, and for a high Subjective Deviation, a factor closer to 0 – i.e., a large penalty.

$$SD_{prod} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{1 + SubjDev(i)}} \quad (17)$$

In eq. 17, the N^{th} root is taken to normalize the overall product, yielding *SDprod* scores between 0 and 1.

Expected Subjective Deviation – the 'normal' case

What Subjective Deviation could we expect from a hypothetical "average subject" ? In other words: Given estimates of the *common loudness* and *subjective loudness* for a set of sounds, derived from a loudness assessment experiment with real subjects. Then imagine that the extra "average subject" independently provided his own loudness assessments of the same sounds. Suppose the resulting loudness levels would have the very same between-subject disagreement properties as those of the real subjects. Now, if we regarded this "average subject" as a *loudness model*, then what would that model's Subjective Deviation be?

To answer this question, let us make two assumptions: 1) that the spread in the *subjective loudness* levels is the same for all sound segments, and 2) that the *subjective loudness* levels belong to a normal distribution with the variance σ^2, and with the *common loudness* as mean. These assumption could be expressed as -

$$\forall i : \mathrm{var}(SegmentLevel_S(i)) = \sigma^2 \wedge \mathrm{mean}(SegmentLevel_S(i)) = SegmentLevel(i) \quad (18)$$

The Subjective Deviation of the "average subject"s predictions would then statistically be as follows -

$$SubjDev_{normal}(i) = \frac{\left|N(SegmentLevel(i), \sigma) - SegmentLevel(i)\right|}{1.349 \cdot \sigma} \quad (19)$$

as the $InterQuartileRange = 1.349 \cdot \sigma$, for any normal distribution with standard deviation σ.

From eq. 19 we can calculate the *expected SDmean* and *SDprod* values, for the archetypical subject, as:

$$SDmean = 0.591 \text{ and } SDprod = 0.652 \quad (20)$$

Notice that these expected Subjective Deviation scores are independent of the specific between-listener disagreement in the hypothetical experiment – they are independent of σ. Although the two assumptions, underlying these scores (eq. 18), are not satisfied for the two actual reference data sets, they do not seem very unreasonable. The scores in eq. 20 could be used as a guideline to judge whether a model is more or less accurate than an "average" test subject would be.

5.4 Bootstrap resampling used to compute the confidence intervals

In section 5.3 we considered a set of statistics that could be used to evaluate the loudness models, and in section 6 the values of these statistics are presented. In order to compare the performance of evaluated loudness models, it would also be relevant to assess the amount of *uncertainty* in each of the statistics.

Confidence intervals are commonly used to express a range of values within which the true parameter value will be with a certain probability (e.g. [92]). The computation of the confidence intervals associated with a statistic requires knowledge of the sampling distribution of the statistic. The sampling distribution can in some cases be derived via a theoretical analysis, or alternatively, it may be described empirically. In the latter case, we could consider the variations of the statistic, if numerous samples from the unknown distribution were available. In our case, however, this solution would require many repetitions of the whole listening experiment producing the reference data, simply to observe the variability in the resulting model-evaluation statistics. Because we are unable to sample repeatedly from the 'population' (the unknown true distribution), we can instead sample repeatedly from our original sample, which is itself an estimate of the population. This method of obtaining 'new' samples is known as *resampling* (e.g. [93, 94]).

The **bootstrap** is a resampling technique invented by Bradley Efron and further developed in [95]. The bootstrap can be considered a general technique for assessing uncertainty in estimation procedures, in which computer simulation through data resampling replaces mathematical analysis [96]. The idea of the bootstrap is essentially to assume that our sample *is* the population, and then use the uncertainties observed in sampling from the sample to estimate the uncertainties of sampling from the population (sort of like "pulling oneself up by one's bootstraps"). In a sense, the bootstrap resampling does what the experimenter would have done: repeat the experiment and observe the variations of the results. The empirical description of our statistic's sampling distribution is obtained by re-computation of the statistic using random (re-)samples from the original data. The resampling is done *with replacement* so that any data point can be sampled multiple times. A thousand repetitions or more may be required to obtain a good estimate of the sampling distribution of the statistic.

The use of bootstrap confidence intervals is not uncontroversial [97], and the procedure is not accurate in all situations [98]. On the other hand, any new statistical method will tend to take decades to move from specialized literature into the statistical analyses of the mainstream experimenters [83] (ch.18).

In order to compute the confidence intervals for our non-standard statistical parameters, such as the product and mean Subjective Deviation, we have employed the bootstrap method. There are several different ways of computing confidence intervals using bootstrap resampling. The method used here is the called the "hybrid method" [97] or "the root method" [98]. The hybrid method is relatively simple and reportedly robust. For simplicity, the bootstrap was used to compute the confidence intervals for *all* the statistics to evaluate the loudness models (next section), even though some of them might alternatively have been computed using parametric statistical procedures.

6 COMPARATIVE STUDY: PERFORMANCE OF THE LOUDNESS MODELS

This section presents the evaluation of the loudness models described in sect. 2 and 3. Each model is tested against the two subjective reference data sets summarized in sect. 4, using the measures presented in sect. 5 to display and quantify how well a model is able to predict the subjective long-term loudness of the sound segments.

6.1 Sub-evaluation of model variants

The PPM and Zwicker&Fastl models were each presented in three different variants, corresponding to the three tested percentile-statistics used in the calculation of the long-term loudness estimate. The best of these variants is identified in this sub-section, to reduce the number of different models being compared. The interpretation of the actual scores is in section 6.4.

6.1.1 PPM model variants

Table 3 shows that the PPM(50%) variant of the PPM-based loudness model achieves the highest accuracy on both reference data sets. This is consistent with the findings of Klar and Spikofski [9], who found the 50th percentile of a PPM-type measurement to achieve the highest rank-correlation with subjective ratings of loudness. The difference between the PPM(50%) and

the PPM(75%) variant is less pronounced on the McGill data set than on the TC set.

Model name	Ref. data	SDmean	SDprod	AAE	RMSE	P95AE	R
PPM50	McGill	1.332	0.480	0.972	1.229	2.352	0.687
PPM75	McGill	1.390	0.478	0.972	1.253	2.342	0.669
PPM95	McGill	1.791	0.419	1.252	1.554	3.137	0.554
PPM50	TC	0.875	0.579	1.123	1.542	2.695	0.437
PPM75	TC	1.062	0.539	1.356	1.854	3.415	0.287
PPM95	TC	1.369	0.477	1.722	2.273	4.335	0.131

Table 3. Evaluation of PPM+percentile model variants. "PPM50" means PPM(50%), etc.

6.1.2 Zwicker&Fastl model variants

The Zwicker&Fastl(95%) variant of the Zwicker&Fastl loudness model achieves the highest accuracy (Table 4) – its RMSE is around 0.5 dB lower than for the two other variants, on both reference data sets. This result is consistent with the N_5 loudness prescribed by Zwicker and Fastl (see section 2.4.1).

Model name	Ref. data	SDmean	SDprod	AAE	RMSE	P95AE	R
Z&F50	McGill	3.430	0.279	2.585	3.192	5.786	0.115
Z&F75	McGill	2.858	0.311	2.099	2.583	4.761	0.220
Z&F95	McGill	2.324	0.356	1.681	2.079	3.599	0.387
Z&F50	TC	1.712	0.442	2.215	3.141	5.530	0.093
Z&F75	TC	1.356	0.472	1.570	2.027	3.604	0.048
Z&F95	TC	1.020	0.542	1.224	1.462	2.918	0.068

Table 4. Evaluation of Zwicker&Fastl model variants. "Z&F50" means Zwicker&Fastl(50%), etc.

6.2 The Absolute Error

The absolute residual error, in dB, for each model's loudness predictions is shown in Figure 4. A histogram together with a boxplot displays the *distribution* of absolute error. (An explanation of the *centered vertical histogram* is provided in the appendix.) We find that this visual display of the data is a good compromise between detail and overview. In both plots the models are ordered according to the trimmed mean of the absolute error.

For the best performing 5 or 6 models, the absolute error is below 1.2 dB for 75% of the sound segments in both the McGill and TC collections. For this group of models, the maximum absolute error is above 3 dB only for a small number of segments in the McGill set. The distribution of error is "chunky with a thin tip", for both the LARM and HEIMDAL models. The L_{eq} models based on the linear, RLB-, and C-weightings, have a similar error distribution shape, though with more errors in the 1-2 dB range. Each of the A-, D-, and M-weighed L_{eq} models, and the Zwicker models, make several prediction errors greater than 5 dB. The L_{eq}(M) does a particularly poor job of predicting the loudness levels of the (music dominated) McGill data set, with 25% of the errors greater than 4 dB, and 12% greater than 5 dB.

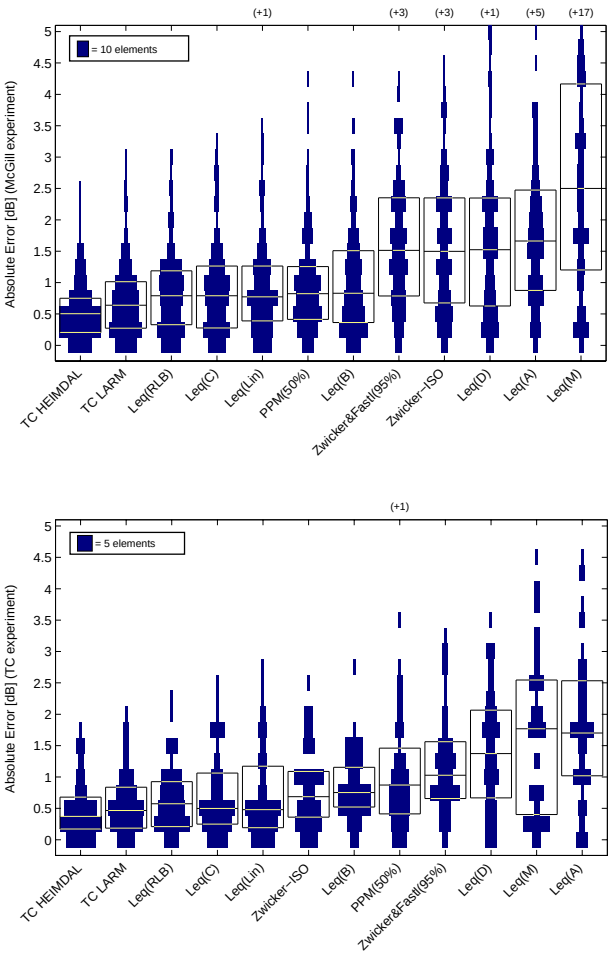

Figure 4. Each centred, vertical histogram shows the distribution of *absolute error* (in dB) for each of the evaluated loudness models. The subjective reference data is (*top*) the McGill set, and (*bottom*) the TC set.

When considering the plots of absolute error of the model predictions one must keep in mind that the subjective reference data are not perfect – any data resulting from a listening experiment will have an uncertainty. For the McGill and TC experiments, this uncertainty was quantified as the standard error of the *common loudness* estimates of the *SegmentLevel*. This value was 0.18 dB for the McGill experiment, and 0.43

dB for the TC experiments. Any error in the predictions made by the models, which is smaller than these standard errors of the reference data, should in principle not be considered an error at all.

6.3 The Subjective Deviation

A centered vertical histogram of the Subjective Deviation for each model is displayed in Figure 5. In both plots, the models are ordered according to the trimmed mean of their respective Subjective Deviation.

The Y-axis and its scale is the same for the two plots corresponding to the two reference data sets. Yet all the models achieve a considerably lower (i.e. better) Subjective Deviation with the TC data set than with the McGill data set as reference. This disparity can have two causes: 1) the McGill collection contained more segments, from a broader range of genres, than the TC collection, and was thus a more *difficult* test for the models, 2) the results from the McGill experiment were based on more adjustments than the TC experiment, and hence the within-subject inconsistency was better suppressed, yielding a smaller spread in the *subjective loudness*. The distribution of absolute error also showed a difference between the two data sets, but to a much lesser degree than reflected in the Subjective Deviation measure. This indicates that the difference is a combination of the two above causes: both the segment collection and the *subjective loudness* spread contributes to the different evaluation of the models.

Aside from the difference between the reference data sets, the overall picture is similar to the distribution of absolute error. In particular, the HEIMDAL and LARM models again make the best predictions, and the $L_{eq}(A)$ and $L_{eq}(M)$ produce the poorest predictions, of the 12 evaluated models.

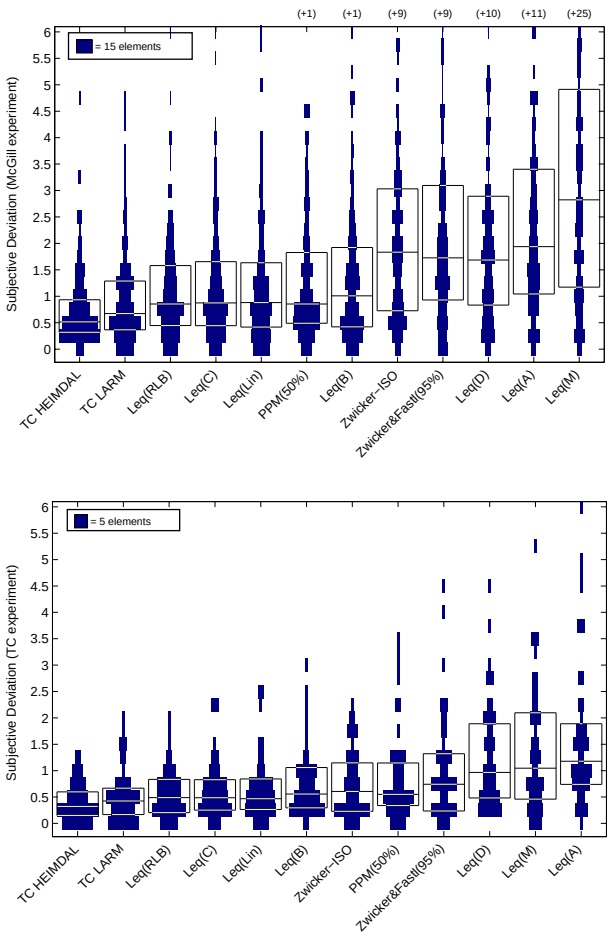

Figure 5. Each centred, vertical histogram shows the distribution of Subjective Deviation for each of the evaluated loudness models. The subjective reference data is (*top*) the McGill set, and (*bottom*) the TC set.

6.4 Model comparison by evaluation measures

Rather than characterizing the performance visually, as in the previous sections, various scalar measures or scores could be considered, as described in section 5. Table 5 presents the five different evaluation measures for the 12 loudness models, with the two different reference data sets. The theoretical optimal and worst values, for the five different measures, are shown in Table 6.

The correlation coefficient could not be computed for the L_{eq}(B), because this loudness model was used to level-normalize the sound segments in the listening

Model name	Ref. data	SDmean	SDprod	AAE	RMSE	P95AE	R
Leq(A)	McGill	2.598	0.328	1.854	2.287	3.826	-0.274
Leq(B)	McGill	1.397	0.473	1.020	1.324	2.636	n/a
Leq(C)	McGill	1.201	0.503	0.918	1.181	2.494	0.531
Leq(D)	McGill	2.415	0.353	1.724	2.159	4.343	0.165
Leq(Lin)	McGill	1.240	0.499	0.980	1.327	2.537	0.485
Leq(M)	McGill	3.861	0.266	2.709	3.358	5.769	0.250
Leq(RLB)	McGill	1.124	0.515	0.855	1.084	2.173	0.581
PPM(50%)	McGill	1.332	0.480	0.972	1.229	2.352	0.687
TC HEIMDAL	McGill	0.769	0.601	0.578	0.748	1.534	0.825
TC LARM	McGill	0.973	0.551	0.744	0.964	1.983	0.707
Zwicker&Fastl(95%)	McGill	2.324	0.356	1.681	2.079	3.599	0.387
Zwicker-ISO	McGill	2.397	0.358	1.711	2.181	3.845	0.417
Leq(A)	TC	1.626	0.425	1.781	2.124	4.126	-0.470
Leq(B)	TC	0.751	0.603	0.840	0.990	1.589	n/a
Leq(C)	TC	0.632	0.640	0.722	0.974	1.954	0.619
Leq(D)	TC	1.248	0.485	1.423	1.687	3.001	0.072
Leq(Lin)	TC	0.655	0.637	0.767	1.063	2.155	0.633
Leq(M)	TC	1.341	0.480	1.684	2.097	3.832	0.249
Leq(RLB)	TC	0.576	0.659	0.667	0.843	1.579	0.638
PPM(50%)	TC	0.875	0.579	1.123	1.542	2.695	0.437
TC HEIMDAL	TC	0.413	0.727	0.520	0.688	1.496	0.815
TC LARM	TC	0.523	0.686	0.608	0.776	1.641	0.634
Zwicker&Fastl(95%)	TC	1.020	0.542	1.224	1.462	2.918	0.068
Zwicker-ISO	TC	0.753	0.604	0.839	1.044	1.954	0.443

Table 5. The 12 models evaluated using the 5 different statistics, with 2 different reference data sets.

experiments, hence the variance approaches zero. This level-normalization is also the reason why the L_{eq}(A) can have a *negative* correlation – this model does worse than doing nothing at all (with level-normalized sound segments).

The HEIMDAL and LARM model parameters were optimized using the McGill data set. Hence, when tested on this data set, a bias towards a better performance could be expected, although cross-validation was employed to counteract this problem (section 3.2). For these two models, the performance relative to the other models is nearly the same for both their training set (McGill data) and validation set (TC data). This indicates that the potential bias in their evaluation is probably quite small.

	SDmean	SDprod	AAE	RMSE	P95AE	R
Optimum (theoretical)	0.0	1.0	0.0	0.0	0.0	1.0
Worst (theoretical)	∞	0.0	∞	∞	∞	-1.0

Table 6. Theoretical optimal and worst evaluations.

Figure 6 and Figure 7 show the AAE, R, SDmean, and SDprod measures[7] together with their *95% confidence intervals* (CI). In all four plots the models are ordered according to their measurement for the McGill experiment (indicated by the filled circle). The CIs of the evaluation measures were computed using bootstrap resampling (section 5.4). Specifically, the CIs quantify the uncertainty caused by the sampling error associated with the particular selection of sound segments, on which the models are evaluated. The larger the selection of sound segments, from the same genres, the more certain the results of the evaluation would be.

As the bootstrap resampling is inherently a stochastic process, it will not produce the exact same results for each run. However, the greater the number of resamples, the more consistent the results will be. In this study, 5000 resampling iterations were used in computing each individual CI. Consequently the first 3 digits of the CI limits were roughly constant between runs.

Under the normal-distribution assumption, the *expected SDmean* = 0.59 was calculated (section 5.3.2). The *SDmean* plot in Figure 7 indicates that only the HEIMDAL model could perform close to the hypothetical "average-subject", in the McGill experiment. In the TC experiment, the 5 best-performing models obtained an *SDmean* evaluation at or above the level of the hypothetical "average-subject".

[7] The RMSE measure was omitted here, due to its similarity to the AAE measure.

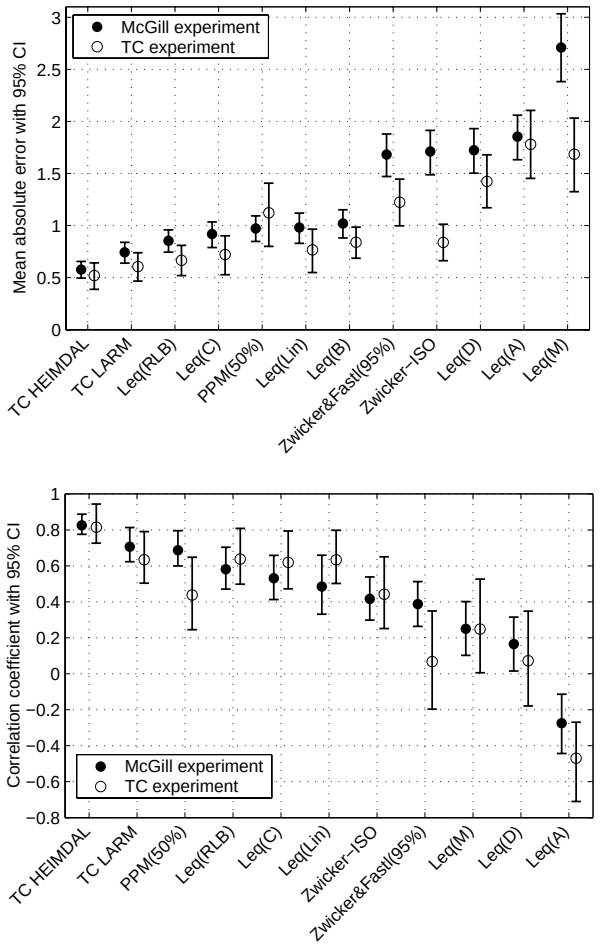

Figure 6. The mean absolute error (*top*) and correlation coefficient (*bottom*) for the evaluated loudness models. For each measurement, the 95% confidence interval is indicated.

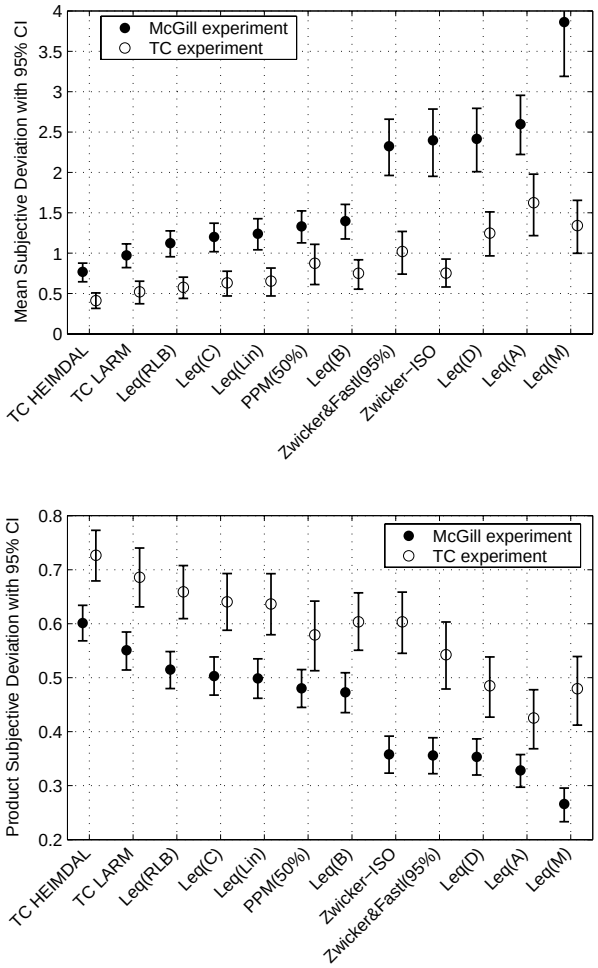

Figure 7. The mean Subjective Deviation (*top*) and product subject deviation (*bottom*) for the evaluated loudness models. For each measurement, the 95% confidence interval is indicated.

Table 7 and Table 8 show the ranking of the models, as determined by four different evaluation measures. The SDmean, SDprod, and AAE measures produce virtually identical ranking of the models, given the same reference data set. The ranking is also nearly – but not quite – the same across the two data sets. The PPM(50%) model is ranked relatively higher with the McGill data set, and the Zwicker-ISO model is ranked relatively higher with the TC data set.

SDmean	–SDprod	AAE	–R
TC HEIMDAL	TC HEIMDAL	TC HEIMDAL	TC HEIMDAL
TC LARM	TC LARM	TC LARM	TC LARM
Leq(RLB)	Leq(RLB)	Leq(RLB)	PPM(50%)
Leq(C)	Leq(C)	Leq(C)	Leq(RLB)
Leq(Lin)	Leq(Lin)	PPM(50%)	Leq(C)
PPM(50%)	PPM(50%)	Leq(Lin)	Leq(Lin)
Leq(B)	Leq(B)	Leq(B)	Zwicker-ISO
Z&F (95%)	Zwicker-ISO	Z&F (95%)	Z&F (95%)
Zwicker-ISO	Z&F (95%)	Zwicker-ISO	Leq(M)
Leq(D)	Leq(D)	Leq(D)	Leq(D)
Leq(A)	Leq(A)	Leq(A)	Leq(A)
Leq(M)	Leq(M)	Leq(M)	

Table 7. Ranking of the models, evaluated with four different statistical measures; best-performing models at the top. McGill data set as reference.

SDmean	–SDprod	AAE	–R
TC HEIMDAL	TC HEIMDAL	TC HEIMDAL	TC HEIMDAL
TC LARM	TC LARM	TC LARM	Leq(RLB)
Leq(RLB)	Leq(RLB)	Leq(RLB)	TC LARM
Leq(C)	Leq(C)	Leq(C)	Leq(Lin)
Leq(Lin)	Leq(Lin)	Leq(Lin)	Leq(C)
Leq(B)	Zwicker-ISO	Zwicker-ISO	Zwicker-ISO
Zwicker-ISO	Leq(B)	Leq(B)	PPM(50%)
PPM(50%)	PPM(50%)	PPM(50%)	Leq(M)
Z&F(95%)	Z&F(95%)	Z&F(95%)	Leq(D)
Leq(D)	Leq(D)	Leq(D)	Z&F(95%)
Leq(M)	Leq(M)	Leq(M)	Leq(A)
Leq(A)	Leq(A)	Leq(A)	

Table 8. Ranking of the models, evaluated with four different statistical measures; best-performing models at the top. TC data set as reference.

6.5 Overall performance of the loudness models

The evaluation measurements seem to suggest a grouping of the loudness models as presented in Table 9. *Class 1* contains the models achieving the best overall evaluation, and *Class 4* contains the models with the worst performance.

Performance class	Models (best-in-class listed first)	Better than class
Class 1	TC HEIMDAL, TC LARM	Class 3, 4
Class 2	Leq(RLB), Leq(C), Leq(Lin)	Class 4
Class 3	Leq(B), PPM(50%), Zwicker-ISO, Zwicker&Fastl(95%)	(none)
Class 4	Leq(D), Leq(A), Leq(M)	(none)

Table 9. Classes of loudness models, based on the overall evaluation. Column 3 refers to a significance level, α=0.05.

The two new models, HEIMDAL and LARM, were both in Class 1, as they obtained a superior accuracy with both the absolute error and the Subjective Deviation measures. The HEIMDAL performed better than any of the other evaluated loudness models, with an average error of only 0.58 and 0.52 dB, for the McGill and TC data sets. HEIMDAL's worst case error, as measured by the *P95AE*, was only 1.5 dB (both data sets). A bias towards positive evaluation could be expected because both Class 1 models were optimized on the McGill data set (see section 3.2.1), but as the models obtained a similar evaluated on the TC data set, we expect this bias to be quite small. Judging from the 95% confidence intervals, the Class 1 models are more accurate than the Class 3 and 4 models, with statistical significance, for both reference data sets (although the CI of LARM overlap slightly with L_{eq}(B) and Zwicker-ISO for the TC set, Class 1 and 3 are otherwise well-separated).

The three L_{eq} models in Class 2 are significantly better than Class 4, for both data sets. The mutual ranking of these models is invariant to different reference data sets and evaluation measures. The only difference between the Class 2 models is their frequency weighting: RLB, C, and none. These models have an *RMSE* of 1.08, 1.18, and 1.33 dB (McGill data), respectively. In comparison to the Class 4 L_{eq} models, with an *RMSE* of 2.2 to 3.4 dB (McGill data), we note that the specific frequency weighting is quite important, and also that the unweighted L_{eq} performs better than 4 out of 6 of the evaluated weightings. The worst case error of the Class 2 models is a *P95AE* of 1.6-2.2 dB (TC data), and 2.2-2.5 dB (McGill data).

Class 3 contains the two variants of the multi-band Zwicker loudness model. Their performance is quite similar, especially for the McGill data set, suggesting that their implementation-differences (section 2.4.1) are not of crucial importance. The extra computational complexity of these models does apparently not improve the accuracy beyond Class 3, when the stimuli belongs to music/speech genres like the McGill and TC segment collections. The PPM(50%) model is close to Class 2 performance for the McGill data set, but not for the TC data. The PPM(50%) is interesting, however, because it is based on a different principle than L_{eq}, and also because PPM measurement (meters) are already installed in many audio production and broadcast studios.

The three L_{eq} models constituting Class 4 produced the least accurate predictions of the loudness levels. The Class 4 models made prediction errors larger than 4.1 dB for 5% of the sound segments, for one or both reference data sets. In particular, the *P95AE* measure for the L_{eq}(M) was 5.8 dB (McGill data). Interestingly, L_{eq}(M) had a better correlation than L_{eq}(A) and L_{eq}(D), for both data sets, yet made many large errors, both on the absolute and the subjective scales. This inconsistency suggests that the L_{eq}(M) might not be measuring the loudness in dB at all, but rather some other unit requiring an extra mapping to dB.

7 RELATION TO PREVIOUS WORK

The lack of a common set of reference data, and a standard method of evaluation, tends to make the comparison between results of different studies difficult. In sections 5 and 6 we argued that in some cases it is meaningless to compare, for instance, measurements of correlation or standard error across different

experiments. In the following, we try to report the findings of other studies that evaluate loudness models with time-varying sounds, in terms of the measures presented and used in our own model evaluation.

Two listening experiments were performed by Jones and Torick, in order to validate their new Loudness Indicator [2]. The stimuli consisted of 12 sound segments of dynamically processed material, in an experiment with 30 subjects, and 8 segments of unprocessed material, in the other experiment with 11 subjects. All segments were only 1.5 to 2 seconds of duration, and of the total 20 segments, the majority contained speech. The performance of the Loudness Indicator was evaluated by comparing its predictions to the median of the subjects' loudness assessments, i.e. a kind of *common loudness* estimate. The *subjective loudness* was also taken into account, by presenting the semi-quartile range of the loudness assessments. In other words, our *Subjective Deviation* measures are comparable to the properties considered by Jones and Torick. The results of the model evaluation was, in our terms, that an AAE of 0.58 dB, for the processed segments, and 1.25 dB, for the unprocessed material. The worst case error was 2.5 dB, and the error was greater than 1.5 dB for only 3 segments. These measurements might indicate that the Loudness Indicator would be a Class 3, perhaps 2, loudness model. However, the reported error statistics must be somewhat uncertain, as they are based on a only 20 sound segments.

Caric and Guzina conducted a study of the loudness in radio broadcasts, in particular the balance between music and speech [99]. Taped broadcast material was measured with a Hewlett-Packard loudness analyzer, implementing a Zwicker loudness model. The announcer's voice was found to have insufficient loudness, compared to music, but the subjective and the measured loudness were not compared quantitatively. PPM and VU-meter readings of the material were also studied, and were found *not* to be suitable for objective loudness measurement.

In the loudness experiments conducted by Aarts [21, 22], subjects matched the loudness of 6 different loudspeakers using pink noise and music stimuli. The objective was – via this loudness equalization – to eliminate loudness differences as a factor in subsequent tests of subjective quality of the loudspeakers. A listening experiment with 10 subjects was used to produce subjective reference data, against which five L_{eq}-type models, and the ISO 532A and 532B models were compared. The stimuli consisted of the different loudspeakers' timbral colorations of the pink noise, around a level of 80 phon. To evaluate the performance of the 7 loudness models, *Hotelling's* T^2 *test* was used. This multivariate statistical test determines whether the model predictions of the relative loudness levels are significantly different from the subjective assessments. The covariance of the subjective ratings, used in the test, can be considered a type of *subjective loudness* (rather than a *common loudness*) as it takes the subjective variability into account. Only the L_{eq}(B) and the ISO 532B (Zwicker) models were *not* significantly different from the subjective reference. The L_{eq}(A) and L_{eq}(D) were found to deviate the most from the reference. This latter result is consistent with the evaluation presented here, whereas the positive performance of the L_{eq}(B) and Zwicker models, reported by Aarts, is not reproduced with the music/speech segment collections used in our study.

Benjamin considered objective measures of loudness, in relation to typical broadcast material [79]. The principal limitations of weighted-L_{eq}–type models compared to Zwicker type models were reviewed. An experiment was then reported, with 210 sound segments, each of 10 seconds duration, consisting mainly of dialogue. The loudness level of the segments were calculated using the L_{eq}(A) and L_{eq}(B), as well as published implementations of the multi-band models by Zwicker and by Moore & Glasberg. As *no subjective reference data* was available, scatter plots were presented to compare the models to each other (no quantitative comparison was reported). When assuming the Zwicker model as reference, the loudness measurements of L_{eq}(A) were within +/–2.5 phon, and the L_{eq}(B) was thought to be slightly closer to the reference. The Zwicker and the Moore & Glasberg models were reported to function nearly identically, in measuring the relative loudness level.

In Part IV of their extensive study of multi-channel level alignment, Zacharov and Bech investigate the ability of loudness models to predict subjective level calibration [100]. Specifically, they consider "*whether a test signal / metric combination can be found that will provide a perceptually valid level alignment compensation for differences associated with different source locations, distances, directivities, sensitivities and asymmetries associated with the room acoustics.*". This objective is clearly different from our investigation of loudness models predicting the loudness level of

music/speech. However, two properties distinguish Zacharov's investigation from the other loudness model evaluations reported here. First, most other studies (including ours) presume that the relative loudness level, in phon, is approximately the same as a gain adjustment, in dB, within the loudness range of the experiment. Instead of making this assumption, Zacharov et al. employed an optimization procedure in the estimation of the gain adjustment predicted by the model – in essence simulating a loudness matching experiment with the model as 'subject'. Second, an impulse of the reproduction system and room was measured, and applied to the signals, in order to let the loudness models 'hear' precisely what the test subjects heard, rather than the 'dry' electrical signal.

Psychoacoustic listening experiments have often employed *anechoic listening conditions*, in order to eliminate the acoustic influence of a room (section 1.1.4). Most loudness experiments involving music and speech as stimuli, however, have used more realistic listening conditions. This presents a mismatch between the loudness models, using the 'dry' signal as input, and the test subjects, listening to the acoustic reproduction of the stimuli in a (somewhat) reverberant room. In essence, the model must also take the influence of the room into account, when predicting the subjects' loudness assessments. The consequence of this disparity could be subject of further investigation.

In order to compare Zacharov and Bech's model evaluation results [100] to our own, we consider their evaluation based on the listeners adjusting the relative level of pink noise played through the individual channels in a surround system, in order for each channel to match the centre channel in perceived loudness. Both the correlation and a regression analysis was used in their model evaluation. The highest correlation ($r=0.2$) was found for the Zwicker and Moore&Glasberg models, both in diffuse and free field modes. The best performing L_{eq}-type model was L_{eq}(A) ($r=0.21$), and the worst L_{eq}(C) ($r=0.14$). Nevertheless, Zacharov and Bech ended up recommending a high-pass filtered test signal that could be used in combination with either of the L_{eq} or multi-band models, as suitable ($r=0.83$) for their calibration task.

Klar and Spikofski report the problem of loudness differences between – and within – radio and TV programmes, and propose some solutions for these problems, based on leveling recommendations and the introduction of a loudness meter [9, 10]. Six different loudness models, based on signal power and signal level (PPM), were evaluated. A long-term loudness prediction was calculated via the envelopes recorded from the loudness meters using a histogram function (see section 2.3.1). A total of 56 sound segments, of 15 seconds duration, were recorded from Digital Satellite Radio. The subjective reference data was formed out of averages of loudness assessments, after mapping from an ordinal scale onto an interval scale [65]. In the evaluation, the rank-order correlation (Spearman's rho) was calculated, between the levels predicted loudness models and the subjective reference data. The two models with the highest correlation (78%) were based on PPM meters with 10 ms integration time and the 50th percentile of their level histogram.

Moore, Glasberg and Stone investigated various effects of amplitude compression applied to speech segments [25]. A listening experiment was conducted with 6 subjects. A total of 18 speech segments of 2.1 seconds duration were extracted. Pairs of stimuli were then formed by matching the uncompressed and a compressed version of a segment – three different compression ratios were employed. The experiment was performed around three levels: 50, 65, and 80 dB SPL. Moore et al. found that, for a fixed RMS level, the compressed speech segments were significantly louder than the uncompressed versions, by up to 3 dB (for a high compression ratio). Predictions of loudness levels for the same stimuli were then computed by the multi-band loudness model of Glasberg and Moore [35]. Specifically, the model computes an instantaneous loudness, from which a (time-varying) short-term loudness is calculated by a temporal integration. The long-term loudness estimate is then calculated by an average over a temporal integration of the short-term loudness envelope. The relative levels thus predicted by the loudness model were consistent with the subjective reference data. The absolute errors were typically below 0.5 dB. In comparison, the McGill segment collection (see section 4.2) also consist of both uncompressed and compressed (mastered) speech and music segments. However, the McGill collection only contain one version of each segment, as opposed to the stimuli used by Moore et al. The mean absolute error was around 0.7 dB and 0.9 dB for the Class 1 and 2 models, respectively, for this considerably more varied segment collection.

Soulodre and Norcross evaluated the performance of 7 L_{eq}-type loudness models [55]. The models were based on the A-, B-, and C-weighting curves, together with

two previously published frequency weightings, and the unweighted L_{eq}. In addition, the Revised Low-frequency B-weighting (RLB), was introduced (see section 2.2.1). A total of 48 sound segments, primarily speech and music, were collected from various broadcast sources. Based on a loudness-matching experiment, the subjective reference data consisted of the adjusted gain level in dB, for each segment, averaged across the 25 subjects [41]. To evaluate the models, Soulodre introduced the Loudness Performance Index, LPI (see section 5.3), and also measured the correlation coefficient, the RMSE, and the maximum absolute error. In addition, a scatter plot of each model's predictions against the reference data was used as a visual supplement to the numerical evaluation measures.

The segment collection used by Soulodre et al. is comparable in size and contents to the TC collection presented here. Furthermore, 5 of the L_{eq} models recur in both studies. The mutual ranking of those 5 models, as found in this study, matches exactly the ranking by Soulodre's correlation and LPI measures, as well as his 'visual rank'. That is, the L_{eq}(RLB) performed the best, followed by L_{eq}(C) and L_{eq}(Lin), with the L_{eq}(A) as the worst. Although it is unclear whether the RLB weighting was fitted to the subjective reference data in [55], our evaluation of the L_{eq}(RLB) with the McGill and TC reference data confirms the superiority of this loudness model over the other tested L_{eq}-based models.

The correlation coefficients in Soulodre's evaluation are between 0.92 and 0.98, for all the models. The contrast to the much greater range of values in our evaluation, demonstrates the arbitrariness of the correlation coefficient (sect. 5.3.1). Soulodre finds an RMSE of the L_{eq}(RLB) of 1.35 dB, compared to only 0.85 and 1.08 dB for the TC and McGill data sets, respectively. Assuming that the segment collection used in Soulodre's study is *not* more difficult than the TC and McGill collections, this discrepancy might be caused by the different experimental method, used by Soulodre et al., leading to a somewhat greater uncertainty in their subjective reference data.

In section 1.2 we described the ongoing study by the SRG-3 within the IRU-R WP6P, concerning the standardization of a new audio metering method. Part of the SRG-3 work has been to evaluate the ability of various loudness meters (i.e., a kind of loudness models) to predict a set of loudness assessments. The SRG-3's subjective reference data was derived from listening experiments at 5 separate sites around the world, involving a total of 96 subjects. The stimuli consisted of 48 sound segments recorded from TV and radio broadcasts, 35 of the segments contained speech. The results of the first round of loudness meter evaluation have been presented in [46, 42]. For this evaluation, 10 loudness meters were submitted by 7 research organizations and private companies. Some of the meters were actual hardware meters, measuring a short-term loudness, whereas others were software implementations, measuring the long-term loudness directly. Numerous error measures were employed in the evaluation: the correlation coefficient, rank-order correlation, RMS-, mean-, and maximum-error, two forms of the LPI, and also the two forms of our Subjective Deviation, *SDmean* and *SDprod* (section 5.3.2). The evaluation also included two loudness models, intended to provide a 'baseline' performance: the L_{eq}(Lin) and L_{eq}(RLB). Yet, these two models obtained a better evaluation than any of the other models, according to most of the error measures.

Note that a version of the HEIMDAL model (preliminary due to time constraints) was submitted to this evaluation within the SRG-3. However, both the structure and the optimization of the HEIMDAL model evaluated here has been substantially improved, compared to the initial version of the model, towards making more accurate and robust estimates of loudness.

It turned out that one of the meter proponents, Dolby Labs, had submitted two additional 'baseline' models to the evaluation: the A- and B-weighted L_{eq}. Consequently, four identified models were evaluated by the SRG-3: A-, B-, and RLB-weighted and unweighted L_{eq}. As these four models were also part of the our evaluation, the two sets of results can be compared. Four evaluation measures can be compared across the different reference data sets: the two measures of absolute error, AAE and RMSE, and the two measures of Subjective Deviation. For both the SRG3, McGill, and TC reference data sets, the mutual ranking of the four L_{eq} models is the same. The average Subjective Deviation (*SDmean*), measured for the 4 models that the 3 data sets have in common, is shown in Figure 8. The SRG3 data was reported in [42] and the data calculated for the McGill and TC data sets was presented in section 6.4.

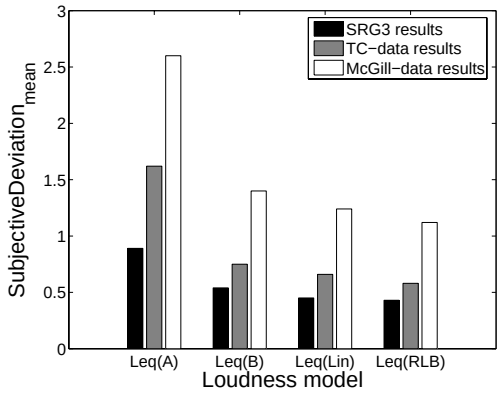

Figure 8. The SDmean measure, reported from the SRG3, and calculated for the McGill and TC data sets.

Figure 8 shows that the ranking of the four L_{eq} models is the same, for the 3 reference data sets. However, the *SDmean* values resulting from the SRG3 evaluation are consistently lower than for TC data set, which again are considerably lower than the *SDmean* for the McGill data set. Recall that the *SDmean* is the average deviation from the *common loudness* estimate, relative to the spread in the *subjective loudness*. Thus the figure indicates that, in the SRG3 study, all the models received a somewhat more positive evaluation, than when tested against either the TC or McGill data sets.

Three different phenomena can cause this consistent difference in Subjective Deviation between the reference data sets:

a. if a collection of sound segments are somehow 'easier' for the models to predict the loudness of, than with some other collection, then the 'easy' reference data set will yield lower *SDmean* measures,
b. a larger between-listener disagreement, in the assessments of each sound segment, will lead to a wider spread in the *subjective loudness* values, and hence lower *SDmean* measures,
c. a larger within-listener inconsistency, in the assessments of each sound segment, will lead to both a wider spread in the *subjective loudness* values, and also to a less certain *common loudness* estimate, and hence lower *SDmean* measures.

As the SRG3 and TC segment collections are comparable in both size and contents, the (a) can not explain the observed difference alone. The (b) could be affecting the SRG3 reference data, because it was based on a number of different test sites, acoustical environments, and subjects. The (c) have probably affected the SRG3 data more than the TC and McGill data, because the latter were based on the balanced pair-matching experimental design [24].

Note that any subjective evaluation of objective measures will inevitably reflect the quality of the objective results *relative to* the quality of the subjective reference data. The Subjective Deviation measures utilized this subjective variability directly, whereas the AAE, RMSE, and correlation measures are merely *biased* by the uncertainty in the reference data.

8 CONCLUSION

In this study, the performance of twelve loudness models was evaluated. Two listening experiments, using speech and music segments as stimuli, provided the subjective reference data. Different statistical measures were employed in the evaluation of the models, so that both the absolute accuracy of the models and the accuracy relative to the between-listener disagreement were measured. We presented a multi-band loudness model (HEIMDAL) based on a novel algorithm, and a single-band model (LARM) based on a combination of two known measurement techniques. The remaining models were all implementations of common or standardized methods, some of which were intended for loudness measurement and others were constructed for measuring levels for various purposes. Eight of them were single-band models based on two different principles: L_{eq} (equivalent sound level) and *PPM* (Peak Program Meter) level measurement. In addition, two variants of the well-established Zwicker multi-band loudness model were evaluated.

A statistical analysis of the responses from the listening experiments produced two kinds of reference data, against which the models were evaluated. Redundancy, incorporated in the experimental design, was used to reduce the influence of the subjects' inconsistency on the results. In the analysis, all adjustments were used for estimating the loudness level of each individual sound segment, to minimize the overall error. Furthermore, two kinds of bias errors were, for each subject, estimated and removed. The statistical analysis was used to compute, for each individual sound segment, estimates of the *subjective loudness*, based on the adjustments of each individual subject, and of the *common loudness*, based on all adjustments obtained from all subjects in the experiment. In other words, the *subjective loudness*, for each sound segment, is the

distribution of loudness level estimates for the individual subjects, and thus the between-listener disagreement – or 'subjective spread' around the *common loudness* level – is characterized. Two experiments were conducted at different sites, with different test subjects and stimuli, resulting in two independent sets of reference data.

To evaluate the loudness models, two different types of error measures were employed: First, a set of measures, based on the absolute error in dB, i.e., the difference between the model prediction and the *common loudness* estimate of the loudness level. This type of measure may be preferable in some cases because their unit is dB; however, they are biased by the 'noise floor' of the underlying subjective reference data. We proposed a second type of measure, the Subjective Deviation, which quantifies the error of the loudness model *relative* to the spread in *subjective loudness* for each segment. Hence, these measures reflect both the accuracy of the model predictions and of the reference data. In particular when comparing results based on different reference data sets, we found it to be essential to include both types of evaluation measures.

The distribution of absolute error or Subjective Deviation was also presented graphically with the *centered, vertical histogram* – a combination of a histogram and a boxplot, showing the quartile error values and the shape of the error distribution, as well as any outliers. We found the centered, vertical histogram to be a more complete and intuitive representation of an evaluated model's performance, than any of the scalar evaluation measures.

A traditional measure of fit, the *correlation coefficient*, was also computed to quantify the deviation of a model's predictions. We found the correlation to depend strongly on the *spread* of the reference data, such that a large spread would lead to a high correlation for all the models, hence compressing their differences. Therefore, the correlation coefficient could be very misleading, if used to compare studies with different reference data sets.

In the field of statistics, the concept of *significance* is of prime importance: if an observed result is not very unlikely to have been caused by chance, the result is not valid. In the case of loudness model evaluation, we should in principle not distinguish between any models whose tested accuracies are not significantly different. We have computed the confidence intervals, for each of the evaluation measures, using a *bootstrap resampling* technique. Thereby we are able to assess the statistical significance of the observed differences between the models. We are not aware of any earlier (loudness modeling) studies using this approach.

Whether a given performance difference between two models is significant will depend on the sample size: the number of sound segments, number of test subjects and loudness assessments. In the practical application of a loudness model, it may make little difference whether the superiority of a given model was found to be statistically significant. It would be more relevant to examine, for instance, the degree of listener's tolerance towards fluctuations in the loudness estimates. In a standardization process, however, where one or more models are declared as superior, is there any reason for *not* using statistical methods to ensure that the observed results are indeed significant, and not merely caused by the specific data sample?

The evaluation measurements seem to suggest a grouping of the loudness models into 4 classes, with *Class 1* containing the models achieving the best overall evaluation, and *Class 4* the worst. Our two new loudness models, HEIMDAL and LARM, constitute *Class 1*. *Class 2* contains the RLB-, C-, and linear-weighted L_{eq} measures, in that order. The L_{eq}(B), a PPM-based measure, and the two Zwicker-based models make up *Class 3*. Finally, the D-, A-, and M-weighted L_{eq} measures, forming *Class 4*, performed remarkably poorly as loudness models. The *Class 4* models all made prediction errors larger than 4.1 dB for 5% of the sound segments, with one or both reference data sets; whereas the HEIMDAL model's errors were just larger than 1.5 dB for 5% of the sound segments, with an average error of only 0.58 and 0.52 dB, for the two data sets.

Somewhat surprisingly, two rather widespread loudness measures, the Zwicker model and the L_{eq}(A), ended up in *Class* 3 and 4, respectively. The Zwicker model, which was originally developed and validated using synthetic and/or stationary signals, could not accurately predict the relative loudness of music and speech segments. The worst performance was observed from the L_{eq}(A) and L_{eq}(M) measures, which are implemented in sound level meters and signal analyzers, and applied for loudness measurements in broadcast and cinema.

9 ACKNOWLEDGEMENTS

We would like to thank René Quesnel and Wieslaw Woszczyk, and the sound engineering students at the McGill University in Montréal, for contributing to this study with expert loudness data.

10 REFERENCES

[1] Bauer, B.B. et al. (1967) "A Loudness-Level Monitor for Broadcasting", IEEE Trans.on Audio and Electroacoustics, vol.AU-15:4, pp.177-182.

[2] Jones, B.L. & Torick, E.L. (1982) "A New Loudness Indicator for Use in Broadcasting", in Proc. of the 71st AES Convention, Montreux.

[3] Stone, M.A., Moore, B.C.J. & Glasberg, B.R. (1997) "A real-time DSP-based loudness meter", in Contributions to Psychological Acoustics.

[4] ISO (1975) "Acoustics. Method for calculating loudness level. International Standard ISO 532 (1.ed.)", International Organisation for Standardisation.

[5] Moore, B.C.J., Peters, R.W. & Glasberg, B.R. (1996) "A revision of Zwicker's loudness model", Acta Acustica, vol.82, pp.335-345.

[6] Brixen, E.B. (2001) "Audio Metering", Broadcast Publishing & DK Audio A/S.

[7] Torick, E.L., Allen, R.G. & Bauer, B.B. (1968) "Automatic Control of Loudness Level", IEEE Transactions on Broadcasting, vol.BC-14:4, pp.143-146.

[8] Vickers, E. (2001) "Automatic Long-term Loudness and Dynamics Matching", in Proc. of the AES 111th Convention, New York.

[9] Klar, S. & Spikofski, G. (2002) "On levelling and loudness problems at television and radio broadcast studios", in Oric. of the AES 112th Convention, Munich.

[10] Spikofski, G. & Klar, S. (2004) "Levelling and Loudness - in radio and television broadcasting", EBU Technical Review, vol.2004:Jan.

[11] Lund, T. (2003) "Loudness Control in Digital Broadcast", in Broadcast Asia 2003.

[12] AES Staff Writer (2004) "How loud is my broadcast?", Journal of the Audio Engineering Society, vol.52:6, pp.662-669.

[13] Moerman, J.P. (2003) "A Proposal for Uniformity in TV-sound", in International Broadcasting Convention, pp.347-355.

[14] Moerman, J.P. (2004) "Loudness in TV Sound", in Proc. of the AES 116th Convention.

[15] Emmett, J.R. (1992) "Programme Loudness Metering and Control", in Proc. of the AES 92nd Convention, Vienna.

[16] Emmett, J. & Girdwood, C. (1994) "Programme Loudness Metering", in AES UK Managing the Bit Budget Conference, pp.92-98.

[17] Emmett, J. & Emmett, J. (2003) "Audio levels - in the new world of digital systems", EBU Technical Review, vol.2003:January.

[18] Katz, B. (2002) "Mastering Audio: The Art and the Science", Oxford: Focal Press.

[19] Nielsen, S.H. & Lund, T. (1999) "Level Control in Digital Mastering", in AES 107th Convention.

[20] Nielsen, S.H. & Lund, T. (2000) "0dBFS+ Levels in Digital Mastering", in AES 109th Convention.

[21] Aarts, R.M. (1991) "Calculation of the loudness of loudspeakers during listening tests", Journal of the Audio Engineering Society, vol.39, pp.27-38.

[22] Aarts, R.M. (1992) "A Comparison of Some Loudness Measures for Loudspeaker Listening Tests", Journal of the Audio Engineering Society, vol.40:3, pp.142-146.

[23] Skovenborg, E., Quesnel, R. & Nielsen, S.H. (2004) "Loudness Assessment of Music and Speech", in Proc. of the AES 116th Convention, Berlin.

[24] Skovenborg, E. & Nielsen, S.H. (2004) "Evaluation of Designs for Loudness-Matching Experiments", in Proc. of the Int.Conf. of Subjective and Objective Assessment of Sound (SOAS), Poznan, Poland.

[25] Moore, B.C.J., Glasberg, B.R. & Stone, M.A. (2003) "Why Are Commercials so Loud? -- Perception and Modeling of the Loudness of Amplitude-Compressed Speech", Journal of the Audio Engineering Society, vol.51:12, pp.1123-1132.

[26] Moore, B.C.J. (1989) "An Introduction to the Psychology of Hearing" (3. ed.), London: Academic Press.

[27] Zwicker, E. & Fastl, H. (1999) "Psychacoustics: Facts and Models" (2. ed.), Springer Series in Information Sciences, 22, Berlin: Springer-Verlag.

[28] Stevens, S.S. (1957) "On the psychophysical law", Psychol.Rev., vol.64, pp.153-181.

[29] Goldstein, E.B. (1989) "Sensation and Perception" (3. ed.), Belmont: Wadsworth Publishing Company.

[30] Appell, J.-E., Hohmann, V. & Kollmeier, B. (2001) "Review of loudness models for normal and hearing-impaired listeners based on the model proposed by Zwicker", Z.Audiol., vol.40:4, pp.140-154.

[31] Hartmann, W.M. (1998) "Signals, Sound, and Sensation", New York: Springer.

[32] ISO (1987) "ISO 226: Acoustics - Normal Equal-Loudness Level Contours", Geneva: International Organization for Standardization.

[33] Zwicker, E., Flottorp, G. & Stevens, S.S. (1957) "Critical bandwidth in loudness summation", J.Acoust.Soc.Am., vol.29:5, pp.548-557.

[34] Zwicker, E. (1977) "Procedure for calculating loudness of temporally variable sounds", J.Acoust.Soc.Am., vol.62:3, pp.675-682.

[35] Glasberg, B.R. & Moore, B.C.J. (2002) "A Model of Loudness Applicable to Time-Varying Sounds", Journal of the Audio Engineering Society, vol.50:5, pp.331-342.

[36] Verhey, J.L. (1999) "Psychoacoustics of spectro-temporal effects in masking and loudness perception", Oldenburg, Germany: BIS Verlag.

[37] Buus, S., Florentine, M. & Poulsen, T. (1997) "Temporal integration of loudness, loudness discrimination, and the form of the loudness function", J.Acoust.Soc.Am., vol.101:2, pp.669-680.

[38] Chalupper, J. & Fastl, H. (2002) "Dynamic loudness model DLM for normal and hearing-impaired listeners", Acta Acustica, vol.88, pp.378-386.

[39] ITU-R (2002) "SRG-3 Status Report (2), September 2002", Document 6P/145-E,

[40] ITU-R (2003) "Subjective assessment of loudness characteristics (Australia)", Document 6P/16-E,

[41] Soulodre, G.A., Lavoie, M.C. & Norcross, S.G. (2003) "The Subjective Loudness of Typical Program Material", in Proc. of the AES 115th Convention.

[42] Soulodre, G.A. (2004) "Evaluation of Objective Loudness Meters", in Proc. of the AES 116th Conv.

[43] Yahoo Groups (2003) "ITU-R reflector for the SRG3 at Yahoo Groups", Internet web site: http://groups.yahoo.com/group/srg3list/.

[44] ITU-R (2003) "Call for submission of audio loudness metering methods (Question 2/6. Working Party 6P, Special Rapporteur Group 3)", International Telecommunication Union, Radiocommunication Study Groups.

[45] CRC (2004) "Advanced Audio Systems Group, Communications Research Centre, Canada", Internet web site: http://www.crc.ca/en/html/aas/home/home.

[46] Lyman, S. (2003) "Meeting Report, Montreal Meeting of ITU-R WP6P SRG3, Aug 5 & 6, 2003, Maison Radio Canada, (Incl. revisions from Sept 16, 2003)", Internet URL: http://groups.yahoo.com/group/srg3list/files/, ITU-R WP6P SRG-3.

[47] ITU-R (2004) "Report of sub-working group 6P-3, meeting on level metering", Document 6P/TEMP/60-E,

[48] Hawkins, H.L. et al. (eds.) (1996) "Auditory Computation", Springer Handbook of Auditory Research, vol.6., New York: Springer.

[49] Greenberg, S. & Slaney, M. (eds.) (2001) "Computational Models of Auditory Function", IOS Press.

[50] Moore, B.C.J., Glasberg, B.R. & Baer, T. (1997) "A Model for the Prediction of Thresholds, Loudness and Partial Loudness", Journal of the Audio Engineering Society, vol.45:4, pp.224-240.

[51] Zwicker, E. (1960) "Ein Verfahren zur Berechnung der Lautstärke (A procedure for calculating loudness)", Acustica, vol.10, pp.304-308.

[52] Dolby Laboratories Inc. (1998) "Model 737 Soundtrack Loudness Meter - Leq(m), User's Manual", Internet web page: http://www.dolby.com/products/Model737/737_3.pdf.

[53] Dolby Laboratories Inc. (2004) "LM100 - Broadcast Loudness Meter", Internet web page: http://dolby.com/products/LM100/.

[54] ISO (2002) "ISO/DIS 21727: Cinematography - Method of measurement of perceived loudness of motion-picture audio material", International Standardisation Organisation.

[55] Soulodre, G.A. & Norcross, S.G. (2003) "Objective Measures of Loudness", in Proc. of the AES 115th Convention.

[56] IEC (1979) "IEC 60651, Sound level meters", International Electrotechnical Commission.

[57] Bauer, B.B. & Torick, E.L. (1966) "Researches in Loudness Measurement", IEEE Trans.on Audio and Electroacoustics, vol.AU-14:3, pp.141-151.

[58] Miljøstyrelsen (The Danish Environmental Protection Agency) (1984) "Ekstern støj fra virksomheder,

vejledning nr. 5/1984 (Environmental Noise From Enterprises).", Copenhagen: Danish Ministry of the Environment.

[59] EU (2002) "DIRECTIVE 2002/49/EC - relating to the assessment and management of environmental noise", The European Communities.

[60] ANSI (1994) "American National Standard: Acoustical Terminology, ANSI S1.1-1994", New York: Acoustical Society of America / American National Standards Institute.

[61] Moore, B.C.J. (1982) "An Introduction to the Psychology of Hearing" (2. ed.), London: Academic Press.

[62] Hellman, R. & Zwicker, E. (1987) "Why can a decrease in dB(A) produce an increase in loudness?", Journal of the Acoustical Society of America, vol.82:5, pp.1700-1705.

[63] CCIR (1986) "Measurement of audio-frequency noise voltage level in sound broadcasting. Recommandation 468-4", International Radio Consultative Committee.

[64] Allen, I. (1997) "Are Movies Too Loud?", in SMPTE Film Conference, Los Angeles.

[65] Spikofski, G. (2000) "Lautstärkemessung im Rundfunk-Sendestudio (Loudness measurement in broadcast studios)", in Tonmeistertagung 21, Hannover, pp.604-618.

[66] European Broadcasting Union (1988) "Sound Quality Assessment Material - Recordings for subjective tests (Audio CD)",

[67] Zwicker, E. (1961) "Subdivision of audible frequency range into critical bands (Frequenz-gruppen)", Journal of the Acoustical Society of America, vol.33:2, pp.248.

[68] Widmann, U., Lippold, R. & Fastl, H. (1998) "A computer program simulating post-masking for applications in sound analysis systems", in Proc. Noise-Con 98, Ypsilanti, Michigan, pp.451-456.

[69] Hansen, K. (1996) "Objective reading of loudness of a sound programme", in AES 100th Convention, Copenhagen.

[70] Akustik Technologie Göttingen (1998) "The si++ module for psychoacoustical interpretation of acoustic signals; Version 3.5", Akustik Technologie Göttingen.

[71] Akustik Technologie Göttingen (2004) "Psychoacoustics (Definitions of the measuring units of the sensory magnitudes)", Internet web page: http://www.akutech.de/mainpage/psychoa.htm.

[72] Widmann, U. (2003) "Krach gemessen - Gehörbezogene Geräuschbewertung", Internet web page: http://www.s4s.de/grundlagen/art.htm.

[73] Zwicker, E., Fastl, H. & Dallmayr, C. (1984) "BASIC program for calculating the loudness of sounds from their 1/3 oct band spectra according to ISO 532 B", Acustica, vol.55, pp.63.

[74] Paulus, E. & Zwicker, E. (1972) "Programme zur automatischen Bestimmung der Lautheit aus Terzpegeln oder Frequenzgruppenpegeln", Acustica, vol.27.

[75] Brüel & Kjær (2004) "Basic Sound Level Meters", Internet web page: http://www.bksv.com/1129.asp

[76] Bishop, C.M. (1995) "Neural Networks for Pattern Recognition", Oxford: Clarendon Press.

[77] Hastie, T., Tibshirani, R. & Friedman, J. (2001) "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer Texts in Statistics, New York: Springer-Verlag.

[78] ITU-R (1997) "BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union Radiocommunication Assembly, ITU-R.

[79] Benjamin, E. (2002) "Comparison of Objective Measures of Loudness Using Audio Program Material", in Presented at the AES 113th Convention.

[80] Orban, R. & Ogonowski, G. (2002) "Orban Optimod-FM 8400 vs. Cutting Edge Omnia6 -- an updated engineering comparison", Internet URL: http://www.orban.com/orban/about/pages/bandwidth/vs_omnia.pdf.

[81] Poulsen, T. (2002) "Psychoacoustic Measuring Methods", Lecture note no. 3108-e, Lyngby, Denmark: Ørsted - DTU, Acoustic Technology.

[82] Griffin Technology (2003) "The PowerMate", Internet web site: http://www.griffintechnology.com/products/powermate

[83] Howell, D.C. (2002) "Statistical Methods for Psychology" (5. ed.), Duxbury.

[84] Trochim, W.M.K. (2004) "Research Methods Knowledge Base", Internet web site: http://www.socialresearchmethods.net/kb/.

[85] Belger, E. (1969) "The Loudness Balance of Audio Broadcast Programs", Journal of the Audio Engineering Society, vol.17:3, pp.282-285.

[86] Riedmiller, J.C., Lyman, S. & Robinson, C. (2003) "Intelligent Program Loudness Measurement and Control: What Satisfies Listeners?", in Proc. AES 115th Conv., New York.

[87] SMPTE (2002) "RP 200: Relative & Absolute Sound Pressure Levels for Motion-Picture Multichannel Sound Systems", SMPTE.

[88] Jolliffe, I.T. (1986) "Principal Component Analysis", New York: Springer-Verlag.

[89] Gnanadesikan, R. (1977) "Methods for Statistical Data Analysis of Multivariate Observations", New York: John Wiley & Sons.

[90] Johnson, R.A. & Wichern, D.W. (1998) "Applied Multivariate Statistical Analysis" (4. ed.), New Jersey: Prentice Hall.

[91] Schmidmer, C., Bitto, R. & Keyhl, M. (2003) "Proposal for Loudness Meter Requirements (submitted to the ITU-R SRG3 reflector, July 2003)",

[92] Siegel, A.F. & Morgan, C.J. (1996) "Statistics and Data Analysis - an introduction" (2. ed.), New York: Wiley.

[93] Simon, J.L. (1997) "Resampling: The new statistics" (2. ed.), URL: http://www.resample.com/content/text/

[94] Yu, C.H. (2003) "Resampling methods: concepts, applications, and justification", Practical Assessment, Research & Evaluation, vol.8:19.

[95] Efron, B. & Tibshirani, R.J. (1993) "An Introduction to the Bootstrap", Monographs on Statistics and Applied Probability, New York: Chapman & Hall.

[96] Zoubir, A.M. & Boashash, B. (1998) "The Bootstrap and its Application in Signal Processing", IEEE Signal Processing Magazine, vol.15:1, pp.56-76.

[97] SAS Institute Inc. (2000) "Jackknife and Bootstrap Analyses", Internet URL: http://ftp.sas.com/techsup/download/stat/jackboot.html

[98] Politis, D.N. (1998) "Computer Intensive Methods in Statistical Analysis", IEEE Signal Processing Magazine, vol.15:1, pp.39-55.

[99] Caric, Z.H. & Guzina, B. (1983) "Loudness Balance of Speech and Music in Radio Broadcasts", in AES 73rd Convention, Preprint 1965.

[100] Zacharov, N. & Bech, S. (2000) "Multichannel level alignment, Part IV: The correlation between physical measures and subjective level calibration", in AES 109th Convention, Los Angeles.

[101] Velleman, P.F. & Hoaglin, D.C. (1981) "Applications, Basics, and Computing of Exploratory Data Analysis", Boston, MA.: Duxbury Press.

11 APPENDIX: THE VERTICAL, CENTERED HISTOGRAM

To evaluate the loudness models visually, a new type of plot was employed – a combination between a box plot and a vertical, centered histogram. An illustrated guide to this visualization device is provided in Figure 9. For each group of data, the centered vertical histogram outlines the distribution of values. This is similar to a normal histogram plot, except that it is displayed vertically, and that its bars are centered. The three horizontal lines that are superimposed the histogram provides the same information as in a *boxplot*: the location of the three *quartiles*, i.e., the 25th, 50th, and 75th percentiles [101]. Thus, the middle horizontal line represents the *median* of the data. The legend in the figure's upper-left corner presents the 'Y-axis' of the histograms, that is, how many elements or data points that a certain length of a bar in the histogram corresponds to.

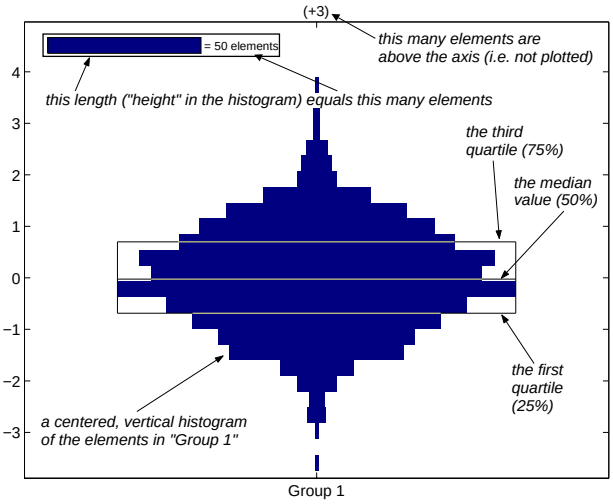

Figure 9. A guide to the *vertical, centered histogram*.