DOI: 10.61782/fa.2023.0600

CONSIDERATIONS FOR THE PERCEPTUAL EVALUATION OF STEADY-STATE AND TIME-VARYING SOUNDS USING PSYCHOACOUSTIC METRICS

Alejandro Osses[1*] **Gil Felix Greco**[2] **Roberto Merino-Martínez**[3]
[1] École normale supérieure, PSL University, CNRS, 75005 Paris, France
[2] Institut für Akustik, Technische Universität Braunschweig, 38106 Braunschweig, Germany
[3] Faculty of Aerospace Engineering, Delft University of Technology, Delft, the Netherlands
ale.a.osses@gmail.com g.felix-greco@tu-braunschweig.de r.merinomartinez@tudelft.nl

ABSTRACT

Sound evaluation based on listening experiments is very costly and time-consuming. Alternatively, the sounds under evaluation can be compared using psychoacoustic metrics, under the hypothesis that sounds that differ by more than one just-noticeable difference (JND) are perceived as significantly different from each other. Well-established psychoacoustic metrics are obtained from models of loudness, sharpness, roughness, fluctuation strength, and tonality. In the present study, we use these psychoacoustic models to analyse a selection of aircraft flyovers and train pass-by sounds, as well as fluctuating sounds obtained from resonances of a rotating flexible tube. In addition to the description of the adopted model implementation of each psychoacoustic metric, our discussion focuses on technical and practical aspects including (1) the dependency of the results on the appropriate sound and model calibration, and (2) how to interpret the results based on reported JNDs from the literature. We reflect on how much we can rely on a specific model implementation considering the status of the metric—standardised (or not), accessibility to algorithms and to their validation results (or not)—and external factors that may influence the obtained estimates.

Keywords: Psychoacoustic metrics, perceptual evaluation, sound quality, steady-state and time-varying sounds

1. INTRODUCTION

Listening experiments have been one of the most traditional ways to compare the perceptual effects between two or more sounds (e.g., [1]). Such a comparison becomes relevant when the test sounds share some acoustic properties. In acoustics, these comparisons are often related to the evaluation of how “salient” a sound source is, assuming that a more salient sound will be more annoying. This is also the case for the acoustic design or characterisation of everyday products (e.g., [2]). The test sounds can also originate from physical simulations, e.g., to implement sound reduction solutions. An example using this approach is given in [3], for the reduction of the interior noise in high-speed trains. A final application example is when recordings are compared to a physical simulation of a given setup and sound source. In this case, the interest is rather fundamental, with the goal of successfully capturing the main properties of the recorded sounds (e.g., [4]).

However, listening experiments are time-consuming, and often require dedicated facilities to be conducted [5]. Also, participants may need to be trained to investigate specific aspects of sound quality, especially when the perceptual aspect to be investigated is not easy to translate into words. As an unfortunate consequence, sound quality evaluations are sometimes informally inspected with very few participants (e.g., [6]), or the evaluations are purely based on qualitative comparisons, without acknowledging the need of perceptual testing (e.g., [7]). To (partly) cope with this problem, psychoacoustic metrics have been developed to ascribe an interpretable number to the sensation produced by sounds (e.g., [8, 9]). These metrics lead to different perceptual scales, which need to be interpreted in terms of the possible range of values and the concept of JND (for a non-exhaustive review, see [10], Chapter 2).

In the current study, we particularly investigate differences in three sets of sounds: (1) aircraft fly-by recordings, (2) train pass-by recordings, and (3) recordings and simulations of a resonating tube. These are sounds that are or have been of interest to each of the authors in recent years. We investigate these sounds in terms of five psychoacoustic metrics—loudness, sharpness, roughness, fluctuation strength, and tonality—for each of which we chose one model implementation. Due to the difficulty to access the information about how these models have been validated, we discuss how, or to what extent, can we interpret the estimates of a specific metric. Although for this study we did not collect experimental data to support our analyses, our sets of sounds and psychoacoustic models were compiled within a MATLAB toolbox that we have called sound quality analysis toolbox (SQAT) [11], with the goal of promoting the reproducibility of results and code accessibility. We pose a strong emphasis on (1) the level conventions used to store, read, and reproduce sounds and how important is to link those level conventions to the "psychoacoustic models" that provide the final perceptual estimates, and on (2) the range of values that have been reported for each metric and the types of sounds that are to be tested, which may or may not have been previously evaluated in the literature.

2. PSYCHOACOUSTIC METRICS

In this section, a description of the selected psychoacoustic metrics is given. The description of these "auditory models" is based on the algorithm implementations available in the SQAT toolbox. To this respect we need to distinguish between the model names and the specific implementations. The model names were adopted from the first author (or standard name) and the year of the corresponding publication, whereas the implementation refers to the specific lines of code that we compiled and included in the toolbox. The implementations are supposed to reflect the original models as closely as possible, but in fact we cannot guarantee that the exact model estimates will be obtained as when using the original codes. However, to grasp the applicability of our model implementations, the toolbox includes a set of validation scripts for each metric.

2.1 Internal level convention

When coding the models for the SQAT toolbox, we standardised all the models to use an internal sound pressure level convention such that the maximum amplitude of a digital waveform (equal to 1 in most digital sound editors) is equivalent to 94 dB SPL. When stored waveforms are read without such normalisation, the maximum amplitude is related to the number of bits used to store the waveforms. For instance, for sounds stored with an amplitude resolution of 16 bits, the full-scale amplitude is ± 32767, corresponding to the normalised ± 1. In this paper, we will always refer to the normalised amplitudes. The definitions of full scale and dB full scale (dB FS) are formalised in several standards, which have been in constant revision [12].

2.2 Description of the psychoacoustic models

In the SQAT toolbox, the selected metrics are obtained from the list of psychoacoustic models indicated in Table 1. In this section, each model is briefly described.

2.2.1 Loudness

Loudness (N) is the subjective correlate of sound pressure level (SPL) and it is expressed in sones [8]. The abbreviation N was adopted for the first time by Fletcher and colleagues [13] to emphasise the difference between loudness numbers, related to the actual perceptual scale, and loudness level which is referenced to equally-loud reference tones, a scale that was later formalised as the phon scale [14]. In general, a doubling of the loudness of a sound leads to a perception of level that is twice as loud as that of the starting level. Loudness is expressed in sones. The code included in the toolbox is a reimplementation for MATLAB of the method for time-varying sounds based on the "Zwicker method" standardised in [15]. In the toolbox, the model is named Loudness_ISO532_1.

2.2.2 Sharpness

Sharpness (S) is a timbre sensation that is used as a measure of the (1) spreading of the frequency components of a sound, and of the (2) presence of high-frequency components relative to the low-frequency components of a sound. The greater the measure the "sharper" the sound. Sharpness is expressed in acum. Sharpness is obtained from a weighted sum of the specific loudness pattern N' of a sound, giving a stronger weighting as the frequency increased. The model was originally proposed by von Bismarck [9] and the toolbox implementation is based on the loudness model (from Sec. 2.2.1), adopting the frequency weighting from the German DIN standard [16]. In the toolbox, the model is named Sharpness_DIN45692.

Table 1: List of psychoacoustic metrics and the corresponding references for their model description and the implementations. For each metric, we report JNDs expressed in percentages.

Metric (abbreviation)	Range (unit)[a]	JND (%)[b]	Model / Implementation adapted from
Loudness (N)	0–120 (sone)	ΔN=7%	[15] / [22]
Sharpness (S)	0–10 (acum)	ΔS=10%	[16] / own
Roughness (R)	0–3.2 (asper)	ΔR=17%	[23] / [17]
Fluctuation strength (F)	0–3 (vacil)	ΔF=10%	[19] / [19]
Tonality (K)	0–1 (t.u.)	ΔK=10%	[21] / own

[a] The maximum values were taken from [18] (Figs. 16.1, 9.1, and 10.2a) for N, S, and F, from [23] (their Fig. 9) for R, and from [21] for K.

[b] The JNDs were taken from [18] (Chapters 11 and 10) for R and F. For N, the JNDs were derived from [24], as described in [10] (his Table 2.1). For S and K we only took a 10% difference as a referential value.

2.2.3 *Roughness*

Roughness (R) is a metric that is typically used for the characterisation of timbre, with a sound being indicated as "rough" depending on the presence of rapid amplitude or frequency modulations in the range of modulation rates f_{mod} between 15 and 300 Hz. The sensation of roughness has a bandpass characteristic with a maximum at a rate f_{mod} of 70 Hz. The metric is expressed in asper. In the toolbox, the model is named Roughness_Daniel1997 and we adapted the code from [17].

2.2.4 *Fluctuation strength*

Fluctuation strength (F) is a metric used to describe slow amplitude or frequency modulations with modulation rates (f_{mod}) below 20 Hz. The sensation of fluctuation strength has a bandpass characteristic with a maximum at an f_{mod} of $\sim$4 Hz ([18], Chapter 10). The metric is expressed in vacil. In the toolbox, the model is named FluctuationStrength_Osses2016 and the code is as described in [19].

2.2.5 *Tonality*

Tonality (K, from the German word *Klanghaftigkeit*) is a metric that evaluates the presence of tonal components in a signal. Tonality ranges from 0 to 1 interpreted as the degree of "tonalness" in the sound, starting from 0 (not having any tonal component at all). The method is based on finding frequency components with salient amplitudes [20] to which tonal and loudness weightings are applied [21]. Tonality is expressed in "tonality units" (t.u.). In the toolbox, the model is named Tonality_Aures1985 and is based on [21]. In the toolbox, however, we use loudness estimates extracted from the more recent model described in Sec. 2.2.1.

Table 2: Acoustic parameters of the selected sounds. T: duration of the sound event (datasets 1 and 2) or of the waveform (dataset 3). SEL=$L_{\text{Aeq},T} - 10 \cdot \log_{10}(T)$.

		Level [dB(A)]			Level [dB]
Sound (type)	T (s)	$L_{\text{Aeq},T}$	$L_{\text{AF,max}}$	SEL	$L_{\text{Zeq},T}$
Flyover2	10.2	85.0	93.4	95.1	88.5
Flyover3	5.4	86.0	91.5	93.3	90.5
Flyover4	10.2	83.2	91.0	93.3	88.2
Flyover6	17.3	76.3	85.5	88.6	83.2
train_01	7.6	80.1	86.8	88.9	80.8
train_11	13.4	79.3	86.8	90.6	80.0
train_15	10.4	79.7	86.6	89.8	79.5
train_20	6.7	82.3	88.1	90.6	82.0
hum-mea	19.9	47.8	50.8	60.7	52.1
hum-syn	5.0	47.5	49.0	54.4	51.8

3. TEST BATTERY

We present three study cases, one provided by each author. The sounds were obtained with different recording setups but for the purpose of this study, we stored them as wav files. We report the sampling frequency and the full-scale convention required to obtain the appropriate reproduction level and, important for this study, to appropriately scale the sounds prior to the use of each psychoacoustic model.

3.1 Aircraft flyover

These recordings were provided by the third author [4]. The sounds were recorded using a sampling rate of 40 kHz, adopting an amplitude convention such that the full-scale amplitude is equivalent to 114 dB SPL. The recording location was the same for all flyovers with a mean overhead height of 66.6 m (between 63.5 and 70.8 m). A summary of the flyover events from the four selected recordings is shown in Table 2. With the a priori knowledge that each recording contained one sound event, an event was automatically identified by looking for the fast A-weighted sound levels that have amplitudes, above the maximum amplitude decreased by the arbitrary level of 25 dB, i.e.,

$$L_{\text{AF}} > L_{\text{AF,max}} - 25 \quad \text{dB(A)} \tag{1}$$

The sound events had durations between 5.4 and 17.3 s with estimated sound exposure levels (SELs) between 88.6 and 95.1 dB(A), as indicated in Table 2. These sounds were chosen to identify and quantify possible perceptual differences, despite their similar SEL values.

3.2 Train pass-by

These recordings were provided by the second author [25]. The sounds were recorded using a sampling rate of 48 kHz, adopting an amplitude convention such that the full-scale amplitude is equivalent to 140.55 dB SPL. This value was obtained from an acoustic calibrator emitting a 1-kHz tone at 94 dB SPL obtained using the same recording chain. The calibration tone had a root-mean-square amplitude of -46.55 dB FS.

The train pass-by recordings were obtained with a microphone located at 7 m from the rail tracks, with all trains driving in the same direction. Four passing-by trains that had similar SELs—between 88.9 and 90.6 dB(A)—were selected as a result of the sound event detection using Eq. 1. This level information is shown in Table 2.

3.3 Corrugated tube

A corrugated tube called hummer resonates when it is rotated at specific speeds eliciting the resonances of the acoustic modes (modes 2 to 6) of the tube. The peculiarity of the hummer sounds is that they are modulated in both amplitude and frequency and that the resulting sounds have a strongly oscillating fundamental frequency (f_0) with very weak harmonics. Two sounds from mode 2 were used, which have an f_0 of around 424 Hz. One of the sounds was obtained from on-site recordings [26] and the other (synthesised) sound was obtained from a numerical model of the hummer [7]. The two sounds had an approximate level of 52 dB SPL, stored at a sampling frequency of 44.1 kHz with a full-scale value of 100 dB SPL.

These recordings and simulated waveforms were provided by the first author and had been previously evaluated in Chapter 2 from [10]. The hummer sounds were considered as quasi-stationary for each rotation period (0.6 s, for this acoustic mode) and although the measured sounds had a duration of 20 s, the simulations were truncated to 5 s. For this reason, the SEL values reported in Table 2 are not relevant for this instrument.

4. RESULTS

The overall results are summarised in Table 3. For each dataset, the sounds that produced the highest percept are indicated by asterisks. When the percept differed by less than one JND across sounds no value is highlighted.

For each dataset, based on the values from Table 3, two sounds were chosen and plotted as a function of time and frequency in Figs. 1 and 2, respectively. In Fig. 1, all psychoacoustic metrics are shown whereas in Fig. 2, only the estimates of specific loudness (N'), specific roughness (R'), and specific fluctuation strength (F') are shown. The metrics of sharpness and tonality provide estimates as a function of time, although in both cases the estimates are obtained from framed-based frequency analyses.

4.1 Aircraft flyover

The two sounds that were selected from dataset 1 had the same event duration (10.2 s) and a similar SEL (see Table 2) but differed in the obtained N and F estimates, as shown in Table 3. The corresponding results are shown in Figs. 1A and 2A. As expected, the maximum loudness was reached at the point when the aircraft was passing in front of the microphone reaching N_{max} values of 128 and 109.1 sones for Flyovers 2 (at t=5.1 s) and 4 (at t=7 s), respectively. The maximum fluctuation strength was estimated for the analysis window that started after the N_{max} values, reaching F_{max} values of 0.84 (at t=5.6 s) and 0.54 vacils (at t=7.2 s), respectively. Those F values are representative of the observation period ending 2 s later, due to the fixed time-window length in the algorithm. For the other metrics, the sounds had similar S estimates, while for roughness, Flyover 4 showed a short peak (R=0.26 asper) right before the maximum loudness, although the overall roughness for both flyovers was relatively low. For tonality, Flyovers 2 and 4 showed sporadic local peaks ($k > 0.1$ t.u.). The frequency analyses from Fig. 2A show that Flyover 2 had more intense components at around 720 Hz (or 6.7 Bark), whereas Flyover 4 has somewhat stronger low-frequency components. From these analyses it can be also observed that the F contributions come from very high frequencies for Flyover 2 (at around 5.9 kHz, or 19.5 Bark) while for Flyover 4 the contribution is stronger for low frequencies.

4.2 Train pass-by

The two sounds that were selected from dataset 2 (train_01 in blue and train_15 in red) are shown in Figs. 1B and Fig. 2B. The selected sounds differed in their loudness, with higher N values for train_01 ($N_{\text{max}} = 77$ sones) than for train_15 ($N_{\text{max}} = 66.2$ sones), despite their similar

Table 3: Estimated psychoacoustic metrics for our test sounds.

Sound	N_{50} / N_{max} (sone)	S_{50} / S_{max} (acum)	R_{50} / R_{max} (asper)	F_{50} / F_{max} (vacil)	K_{10} / K_{max} (t.u.)
Flyover2	46.4 /**128.2***[a]	1.33 / 2.26	0.08 / 0.16	0.18*/ 0.84*	0.07 / 0.12
Flyover3	69.3*/ 114.3	1.58*/ 2.21	0.09 / 0.22	0.13 / 0.69	0.08 / 0.13
Flyover4	43.9 / 109.1	1.37 / 2.33	0.08 / 0.27	0.15 / 0.54	0.08 / 0.14
Flyover6	23.4 / 82.3	1.16 / 2.10	0.07 / 0.55*	0.14 / 0.47	0.08 / 0.14
train_01	48.0 / 77.0	1.69 / 2.02	0.06 / 0.21	0.22*/ 2.50*	0.06 / 0.09
train_11	25.3 / 76.2	1.65 / 2.00	0.07 / 0.23	0.18 / 1.78	0.06 / 0.14
train_15	25.3 / 66.2	1.69*/ 2.23*	0.06 / 0.24	0.19 / 2.06	0.08 / 0.15
train_20	52.3 / 79.9	1.77 / 2.04	0.06 / 0.15	0.24 / 0.73	0.06 / 0.08
hum-mea	2.6*/ 3.6*	0.74*/ 1.49*	0.01 / 0.63*	0.18 / 0.21	0.50 / 0.63
hum-syn	2.0 / 2.5	0.55 / 1.22	0.01 / 0.02	0.24*/ 0.24	0.71*/ 0.77*

(a)The N value in bold is outside the value range indicated in Table 1.

(*)Sound(s) eliciting the maximum percept of the corresponding metric, within datasets. No value is indicated if differences are less than a JND.

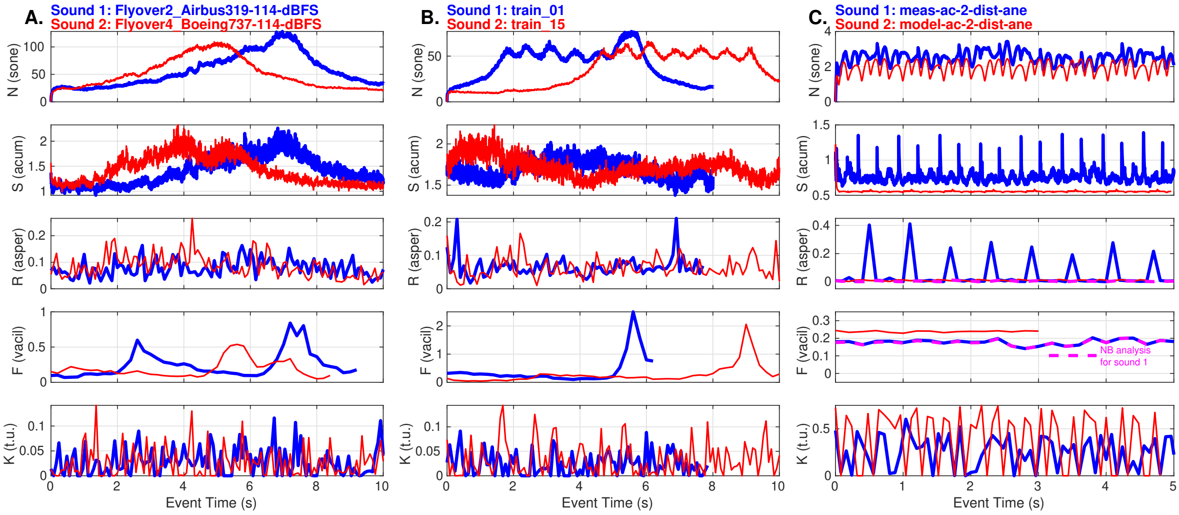

Figure 1: Psychoacoustic metrics as a function of time for a selection of two sounds from datasets 1 (A, left), 2 (B, centre), and 3 (C, right). From top to bottom the metrics are loudness, sharpness, roughness, fluctuation strength, and tonality. The R and F estimates for the measured hummer (C panels, in blue) using a narrow-band analysis are shown in pink dashed lines (see the text for details).

SEL and $L_{\mathrm{AF,max}}$ values (see Table 2). The sounds also had similar S values ($S_{50} = 1.69$ asper) but reached a somewhat higher maximum value for train_15 (S_{max}=2.23 asper). As depicted in Fig. 1B (second row) the higher S_{max} value for train_15 is observed before the train had passed by, indicating that a more refined sound event detection would have not included those maximum values. A similar effect of fluctuation strength was estimated as for the aircraft flyover sounds, with F values being the highest for the time windows including the start of the decaying amplitudes, producing F_{max} values at t=5.6 s (train_01) and 9.0 s (train_15). For the roughness estimates, local maxima were found ($R > 0.2$ asper) before and after the train passed by, which was more visible for train_01 (in blue). In terms of tonality, train_15 had a more prominent tonality, with local peak values K of $\sim$0.1 t.u (or slightly above). The frequency analyses show that the more intense event of train_01 contained stronger low- and high-frequency components, as shown in the N' results (Fig. 2B, top). For roughness, despite the overall low values, the contributions per frequency were different, with a stronger low-frequency contribution for train_01, whereas for train_15 (in red), there was a considerable R' contribution at around 1500 Hz ($\sim$ 11 Bark).

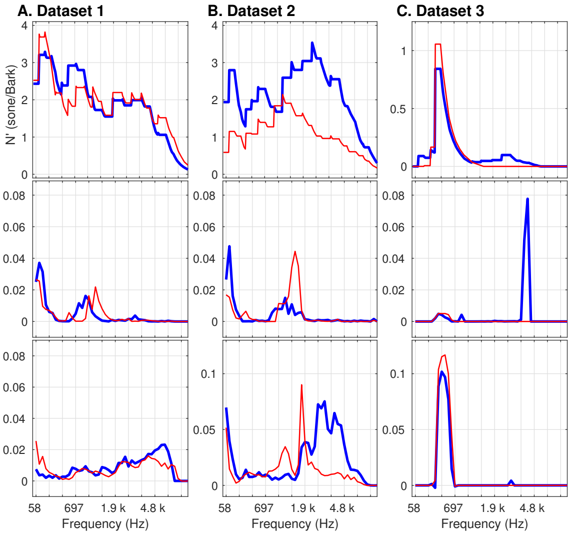

Figure 2: Time-averaged (median) specific loudness (N'), roughness (R'), and fluctuation strength (F') as a function of frequency for the selected sounds from datasets 1, 2, and 3, respectively. The frequency axes are spaced according to the Bark scale [27], between 0.5 Bark (58 Hz) and 23.5 Bark (13.5 kHz). For ease of readability, labels in Hz were added every 1.5 Bark. The sounds and colour codes are the same as in Fig. 1.

4.3 Corrugated tube

The two sounds from dataset 3 are shown in Figs. 1C and 2C. For these two sounds, clear differences were observed from the time signals (Fig. 1C), particularly for sharpness and roughness, where higher estimates were obtained for the measured hummer (in blue). In the case of the R estimates, as depicted in the frequency plot (Fig. 2C, middle), the differences stem from frequencies at $\sim$5 kHz. For sharpness, the differences may originate from off-frequency components (below and above the f_0 of the hummer), as suggested by the frequency distribution of the specific loudness in the Fig. 2C (top) where the measured sound had higher N' values below 278 Hz (2.8 Bark) and above 1117 Hz (9.2 Bark). In terms of tonality, the synthesised sound resulted in a higher K estimate (K_{10}=0.71 for hum-syn; K_{10}=0.50 for hum-mea). These results suggest that there was a strong perceptual influence of frequency components that were not accounted for in the numerical model [7]. These differences are actually related to a measurement noise of the mechanical system on which the hummer was placed [26]. For this dataset, however, since the goal was to compare the fidelity of the synthesised sounds, a more appropriate analysis should have been focused on the similarity between sounds at around the f_0 of the sounds. Such an "ad-hoc" narrowband (NB) analysis is schematised for roughness and fluctuation strength, where the time estimates were reassessed only integrating the specific values between 2.8 and 9.2 Bark. The new R and F estimates are indicated by dashed pink lines in Fig. 1C, where it can be seen that the estimated local R maxima disappeared, while F estimates remained nearly unchanged. This result confirms that the "broadband" roughness (in blue) was indeed related to the measurement noise of the hummer system.

A final analysis that we wanted to include for hummer signals is related to the variability of the perceptual estimates within each rotation period (of 0.6 s). For this analysis, shown in Fig. 3A, we only plotted estimates for loudness derived from the corresponding time signals (Fig. 1C, top). From this analysis, the data points marked as "all" represent a summary of the amplitudes shown in Fig. 1C, whereas the data points marked as "frame" were obtained from median N estimates within 0.6 s segments. In this sense, every 0.6 s section of the measured waveform represents one instance of the measurement. It can be thus seen that there is an intrinsic variability which is related to different but naturally-varying rotations of the hummer. The between-frame values reached values that differ by approximately ±0.2 sones and can thus be in the range of a JND (in this case, a 7% of the mean value is 0.17 sone), leading to a perceptually different value if the judgement is purely based on reading the estimated N values. Hence, when sound source measurements allow it, we recommend evaluating the effect of sound intrinsic variations on the results of a (perceptual) evaluation.

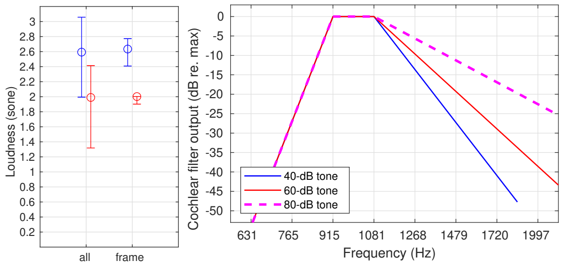

Figure 3: (Left) External variability for hummer signals derived from the loudness time signals, as shown in Fig. 1C ("all") or derived from averages for each rotation period of 0.6 s ("frame"). The "frame" estimates were expected to be constant for the synthesised sound (in red), because the sound was obtained from a numerical model. (Right) Excitation pattern for 1-kHz tones of 40, 60, and 80 dB SPL for the non-linear filter bank used in the R, F, and K models. These patterns are shown to emphasise the level dependency of the filter bank (and thus of the analyses).

5. DISCUSSION AND PERSPECTIVES

5.1 Comparison between sounds

5.1.1 Datasets 1 and 2: Recorded sounds

Datasets 1 and 2 represent examples of sound quality evaluations as they can be required for estimating the annoyance [18] of different sound events or for ranking sounds using a specific perceptual criterion [2]. In general, each psychoacoustic metric represents a different evaluation scale and, thus, these analyses shed light on perceptual differences which are likely to be perceptible in a listening condition. We applied such a rationale to select the two most different sounds from flyovers or pass-by trains, where we were able to quantify their differences using the time and frequency analyses (Figs. 1 and 2).

5.1.2 Dataset 3: Validation of a numerical model

The evaluated synthesised sound came from a physical (numerical) model of the hummer [7]. The results showed that the identified perceptual differences were mostly concentrated outside the range of the fundamental frequency of the hummer resonances. In the measured sounds, that is a frequency region where the engine of the mechanical system used for the recordings was clearly audible [26]. Thus, if the differences stemmed from that frequency range, our sound comparison (Figs. 1 and 2) was not fair, because the synthesised sounds did not contain that noise. From the frequency analyses (Fig. 2C) it seems clear that the roughness differences were concentrated at around 5 kHz and thus, that for the on-frequency components, the roughness between the sounds should be perceptually equivalent. For loudness, although the synthesised hummer estimate was higher and this may be mostly related to the on-frequency components (0.2 sones higher with respect to the measured sounds), it may well be that the high-frequency components (>1500 Hz) were responsible for the significant difference identified for S estimates.

For this dataset, we depicted perceptual estimates for multiple instances (or measurements of 0.6 s), emphasising that recordings coming from real measurements might have a non-negligible effect on the estimated psychoacoustic metrics. This effect is (often) not present in numerical models and should, ideally, be accounted for in any type of sound comparison.

5.2 Level dependency

Most of the psychoacoustic metrics are based on some type of non-linear transformation inspired by underlying hearing processes. An example of such a non-linearity is Terhardt's cochlear filter bank [28] that is used in the selected roughness, fluctuation strength, and tonality algorithms. We illustrate the level dependency of this filter bank in Fig. 3B for a 1-kHz tone adjusted to have 40, 60, and 80 dB SPL (or N=1, 4, and 16 sones) using the perceptually-inspired resolution of 0.5 Bark (of ~80 Hz around 1 kHz). In the case of the R and F algorithms, higher presentation levels of the same type of signal lead to higher estimates due to the shallower slopes towards high frequencies, which increased the correlation between contiguous frequency bands. Although in this example, we deliberately adjusted the same 1-kHz tone to have a specific SPL, the incidental analysis of input sounds with wrong calibration levels may also lead to such differences.

5.3 Reliance on JNDs and model implementations

The analyses presented in this contribution should only be taken as an indication of the perceptual impact of a specific sound. Our evaluation relied not only on the concept of JND but also on the adopted model implementations. The concept of JND used in many sensory modalities [29] depends on the specific task and, therefore, on the sound types being evaluated. The main utility of JNDs is to establish a guideline for the interpretation of the different perceptual (but numerical) scales. Of course, if a computational model is used to quantify a perceptual attribute, an empirical JND might lead to overestimations or underestimations of the scale if the model is inappropriately sensitive or insensitive to that attribute, respectively. In fact, the selected models of loudness, fluctuation strength, and roughness are based on frequency analyses using the Bark scale established in the early sixties [27], which since the eighties has been indicated as underestimating the frequency selectivity of the human hearing [30]. The strong interest in further developing and revising loudness, leading to a recent international standard [15,31], has not been reflected in the study of other metrics, although there have been several research studies attempting to modernise the (fairly) well-accepted old algorithms (e.g., [32, 33]).

5.4 Perspectives

The goal of this contribution was to provide considerations on how to interpret results from published models of five psychoacoustic metrics—loudness, sharpness, roughness, fluctuation strength, and tonality–providing a set of validation scripts and examples to transparently report how the models work for several sound types [11] (see also [34]). It is important to note that although these models will provide a quantitative metric for any sound, these metrics may vary as a function of several factors, such as, presentation level, or for different instances (or takes) of the same sound event type. For this reason the interpretation of results should always be careful and placed in the context, scope, and limits of each adopted algorithm.

6. DATA AVAILABILITY

All figures and analyses from this contribution can be reproduced using the SQAT toolbox, using the script pub_Osses2023c_Forum_Acusticum_SQAT.m. SQAT is an open-source toolbox available on GitHub [11]. The database of sounds can be retrieved from Zenodo [35].

7. ACKNOWLEDGEMENTS

Author AO received funding from French National Research Agency (ANR, project: 17-EURE-0017). Author GFG received funding from the Deutsche Forschungsgemeinschaft (DFG, EXC 2163/1, project: 390881007).

8. REFERENCES

[1] R. Merino-Martínez, R. Pieren, and B. Schäffer, "Holistic approach to wind turbine noise: From blade trailing-edge modifications to annoyance estimation," *Renewable Sustainable Energy Rev.*, vol. 148, p. 111285, 2021.

[2] S. Atamer and M. E. Altinsoy, "Effect of tonality in loudness perception: Vacuum cleaner and shaver examples," *Acoust. Sci. & Tech.*, vol. 41, pp. 369–372, 2020.

[3] K. Qian, Z. Hou, Q. Sun, Y. Gao, D. Sun, and R. Liu, "Evaluation and optimization of sound quality in high-speed trains," *Appl. Acoust.*, vol. 174, p. 107830, 2021.

[4] R. Merino-Martínez, M. Snellen, and D. Simons, "Functional beamforming applied to imaging of flyover noise on landing aircraft," *J. Aircr.*, vol. 53, pp. 1830–1843, 2016.

[5] R. Merino-Martínez, B. von den Hoff, and D. Simons, "Design and acoustic characterization of a psychoacoustic listening facility," in *International Congress on Sound and Vibration*, 2023.

[6] J. Chabassier, A. Chaigne, and P. Joly, "Modeling and simulation of a grand piano," *J. Acoust. Soc. Am.*, vol. 134, pp. 648–65, 2013.

[7] G. Nakiboğlu, O. Rudenko, and A. Hirschberg, "Aeroacoustics of the swinging corrugated tube: Voice of the Dragon.," *J. Acoust. Soc. Am.*, vol. 131, pp. 749–765, 2012.

[8] S. S. Stevens, "A scale for the measurement of a psychological magnitude: Loudness," *Psychol. Rev.*, vol. 43, pp. 405–416, 1936.

[9] G. von Bismarck, "Sharpness as an attribute of the timbre of steady sounds," *Acustica*, vol. 30, pp. 159–172, 1974.

[10] A. Osses, *Prediction of perceptual similarity based on time-domain models of auditory perception*. Ph.D. thesis, Technische Universiteit Eindhoven, 2018.

[11] G. Felix Greco, R. Merino-Martínez, and A. Osses, "SQAT: A sound quality analysis toolbox for MATLAB," 2023. https://github.com/ggrecow/SQAT. Zenodo. 10.5281/zenodo.7934710.

[12] AES, "AES standard method for digital audio engineering–Measurement of digital audio equipment," 2020.

[13] H. Fletcher and W. Munson, "Loudness, its definition, measurement and calculation," *J. Acoust. Soc. Am.*, vol. 5, pp. 82–108, 1933.

[14] D. Robinson, "The relation between the sone and phon scales of loudness," *Acustica*, vol. 3, p. 344, 1953.

[15] ISO, "ISO 532-1:2017. Acoustics. Method for calculating loudness – Part 1: Zwicker method," 2017.

[16] DIN, "DIN 45692–Messtechnische Simulation der Hörempfindung Schärfe," 2009.

[17] J. Schrader, "A MATLAB implementation of a model of auditory roughness," tech. rep., Eindhoven University of Technology, 2002.

[18] H. Fastl and E. Zwicker, *Psychoacoustics, Facts and Models*. Springer Berlin Heidelberg, third ed., 2007.

[19] A. Osses, R. García, and A. Kohlrausch, "Modelling the sensation of fluctuation strength," *Proc. Mtgs. Acoust.*, vol. 28, no. 050005, pp. 1–8, 2016.

[20] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.*, vol. 71, no. 3, pp. 679–688, 1982.

[21] W. Aures, "Berechnungsverfahren für den sensorischen Wohlklang beliebiger Schallsignale," *Acustica*, vol. 59, pp. 130–141, 1985.

[22] D. Cabrera, D. Jimenez, and W. Martens, "Audio and acoustical response analysis environment (AARAE): A tool to support education and research in acoustics," in *Proceedings of Internoise*, (Melbourne, Australia), pp. 1–10, 2014.

[23] P. Daniel and R. Weber, "Psychoacoustical roughness: Implementation of an optimized model," *Acustica - Acta Acustica*, vol. 83, pp. 113–123, 1997.

[24] W. Rabinowitz, *Frequency and intensity resolution in audition*. M.Sc. thesis, Massachusetts Institute of Technology, 1970.

[25] J. Hoffmann, "Comparison of loudness models in the context of railway noise," 2020. Student project, Technische Universität Braunschweig.

[26] M. Hirschberg, O. Rudenko, G. Nakiboğlu, A. Holten, J. Willems, and A. Hirschberg in *Proceedings of SMAC*, (Stockholm), 2013.

[27] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, p. 248, 1961.

[28] E. Terhardt, "Calculating virtual pitch," *Hear. Res.*, vol. 1, pp. 155–182, 1979.

[29] S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, pp. 153–181, 1957.

[30] B. Moore and B. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

[31] ISO, "ISO 532-2:2017. Acoustics. Method for calculating loudness – Part 2: Moore-Glasberg method," 2017.

[32] R. Duisters, "The modelling of auditory roughness for signals with temporally asymmetric envelopes," tech. rep., Eindhoven University of Technology, Eindhoven, 2005.

[33] J. Estreder, G. Piñero, M. de Diego, J. Rämö, and V. Välimäki, "Improved aures tonality metric for complex sounds," *Applied Acoustics*, vol. 204, 2023.

[34] G. Felix Greco, , R. Merino-Martínez, A. Osses, and S. Langer, "SQAT: a MATLAB-based toolbox for quantitative sound quality analysis," 2023. To be presented at *Internoise*, (Chiba, Japan).

[35] A. Osses, G. Felix Greco, and R. Merino-Martínez, "Sound stimuli: Considerations for the perceptual evaluation of steady-state and time-varying sounds using psychoacoustic metrics," 2023. Zenodo. 10.5281/zenodo.7933489.