Comparison of loudness models for artificial and environmental sounds

Jan Rennies[1,2], Jesko L. Verhey[2]

[1] *Fraunhofer IDMT, Hearing, Speech and Audio Technology, Oldenburg, Germany, Email: rns@idmt.fraunhofer.de*

[2] *Institute of Physics, University of Oldenburg, Germany, Email: {jan.rennies, jesko.verhey}@uni-oldenburg.de*

Introduction

The loudness of a sound depends on several parameters, e.g. its level or spectrum. For stationary sounds, loudness models have been developed which can correctly predict equal-loudness level contours or effects such as spectral loudness summation. Most natural sounds, however, are not stationary but have temporally fluctuating envelopes. Such temporal fluctuations can also influence the perceived loudness, which is not accounted for by stationary loudness models. Since the prediction of loudness is relevant for a number of applications, e.g. the fitting of hearing aids or the objective assessment of noise emissions, it is also desirable to accurately model the loudness of time-varying sounds.

In more recent studies, dynamic loudness models have been proposed to also account for such temporal aspects of loudness perception. The work presented in this study investigates the dynamic properties of two elaborate loudness models currently available: the model proposed by Chalupper and Fastl [1] and the one proposed by Glasberg and Moore [2]. While both models are based on the model originally developed by Zwicker [3, 4] and in consequence have a similar structure, there are fundamental differences in their dynamic properties. The aim of this study is to investigate the consequences of these different concepts for the predictions of loudness of time-varying sounds.

Model by Chalupper and Fastl

The structure of the loudness model by Chalupper and Fastl [1] is schematically shown in the left panel of Figure 1. The input time signal is high-pass filtered using a Butterworth filter with a cut-off frequency of 50 Hz to account for the lower limit of the audible frequency range. In the following stage a bank of 24 overlapping critical-band filters is applied. At the output of the filter-bank stage, 24 band-pass filtered time signals are available. For each channel, a temporal window with an equivalent rectangular duration (ERD) of 4 ms is temporally shifted along the signal in steps of 2 ms to compute the short-term root-mean-square (rms) value. The transmission of sound from free-field through outer and middle ear is accounted for by a correction factor in the next stage, resulting in the quantity excitation. The excitation is then transformed to specific loudness in several steps. At first, the quantity main loudness is calculated applying the compressive relation between excitation and loudness and accounting for loudness near threshold in a way very similar to the original model [3, 4]. The exponent describing the compression has a value of 0.23. Then, effects of forward masking are included (the influence of backward masking is neglected). This is achieved in a non-linear stage by appending temporal tails to peaks of the specific loudness. The time constants are chosen according to forward masking experiments and depend on level and duration (see [1] for details). Subsequently, spectral masking is accounted for according to DIN 45631. The resulting specific-loudness-time pattern is then integrated along the frequency dimension. The final loudness is obtained by low-pass filtering the resulting time-dependent variable using a first-order low-pass filter with a cut-off frequency of 8 Hz, which simulates temporal integration of loudness.

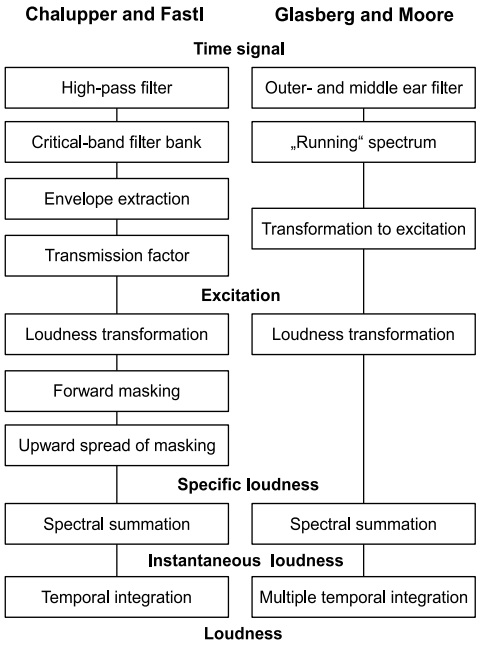

Figure 1: Schematic structure of the loudness model by Chalupper and Fastl [1] (left) and the model by Glasberg and Moore [2] (right).

Model by Glasberg and Moore

The general structure of the loudness model by Glasberg and Moore [2], which is schematically shown in the right panel of Figure 1, is similar to the one by Chalupper and Fastl [1]. However, there are some substantial differences as outlined below. As in the model by Chalupper and Fastl [1], the time signal of the stimulus under consideration is used as input to the model. A fixed filter represents the combined effect of the transfer function from free-field to ear drum and of the transmission through the middle ear. As an intermediate variable, the excitation pattern is calculated from the effective

spectrum reaching the cochlea (i.e. after accounting for the transfer through outer and middle ear). In order to obtain a spectrum which approximates the spectral and temporal resolution of the hearing system for different frequency regions, the filtered time signal is analyzed using six parallel Fast Fourier Transforms (FFTs), each assigned to a different frequency range and calculated with a different analysis window (see [2] for details). The short-term spectra are calculated by shifting the six temporal analysis windows - all aligned at their temporal centres - along the time signal in steps of 1 ms. Each millisecond, the excitation pattern is calculated from the resulting spectra in the same way as in a previous model by the authors [5], accounting for the width of the auditory filters, their level dependence and their variation with centre frequency. Instead of the critical band, Glasberg and Moore use a frequency filtering based on the equivalent rectangular bandwidth (ERB). The excitation pattern is transformed into specific loudness as in the stationary loudness model [5]; as in the model by Chalupper and Fastl [1], compression and the influence of hearing threshold are included in the transformation. The compressive exponent has a value of 0.2. The specific loudness is then summed across frequency. After this stage of the model, loudness is available at a sampling rate of 1 ms, i.e. the same rate at which the spectra and excitation patterns are computed. This loudness is termed "instantaneous" loudness by Glasberg and Moore [2], and interpreted as "an intervening variable which is not available for conscious perception". The instantaneous loudness, which closely follows the temporal envelope of the input signal, is integrated using an attack time constant of about 22 ms and a release time constant of about 50 ms, resulting in a so-called short-term loudness, which is described as "the loudness perceived at any instant" [2]. The short-term loudness is subsequently integrated again in the same way, but with longer time constants for attack and release (99 and 2000 ms, respectively). The resulting long-term loudness is meant to describe loudness sensations that are built rather slowly, e.g. for sounds modulated at a very slow rate.

Loudness of tone bursts

As mentioned in the introduction, loudness is integrated over time. This can be simply illustrated by considering the response of the loudness models to a tone pulse. Figure 2 shows the instantaneous (dashed line) and short-term loudness (solid line) calculated by both models in response to a 1-kHz tone burst at 40 dB SPL and a duration of 200 ms including 10 ms on- and offset ramps. Additionally, for the model by Glasberg and Moore [2], the long-term loudness is shown (dotted line, right panel). It can be observed that the instantaneous loudness closely follows the physical excitation in the model by Glasberg and Moore, while a slower decay is calculated by the one by Chalupper and Fastl. This is due to the forward masking included before the final temporal integration. The latter results in the short-term loudness, which is built up very similarly in both models, while the decay is faster in the model by Chalupper and Fastl [1]. The short-term loudness reaches a value of 1 sone in both models, i.e. the loudness reaches the value of a continuous 1-kHz tone. This is expected, since a duration of 200 ms is already longer than the time constants typically used to describe temporal integration of loudness (e.g. [3, 6]). Thus, for this type of stimuli, the maximum of the short-term loudness is a good estimate of the overall loudness. The long-term loudness, on the other hand, does not reach a stationary value of 1 sone within 200 ms. It should be noted that the long-term loudness was not meant to describe stimuli like short tone burst, but rather to assess the loudness of long stimuli as considered in the following section.

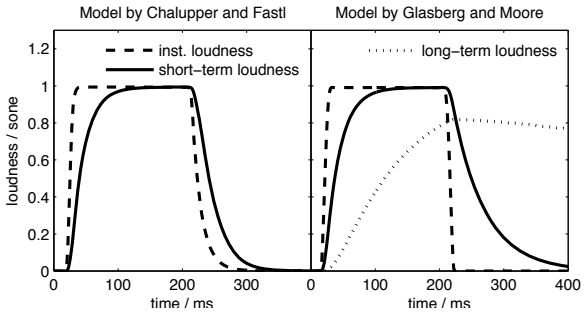

Figure 2: Loudness of a 1-kHz tone burst as a function of time for the model by Chalupper and Fastl (left) and the one by Glasberg and Moore (right).

Loudness of amplitude-modulated tones

Figure 3 shows the level difference between equally loud sinusoidally modulated and unmodulated tones for carrier frequencies of 1 kHz (upper panel) and 4 kHz (lower panel). The modulation depth is $m = 0.5$ in both cases. Model predictions are compared to data from several studies as indicated in the top of the panels. In general, both models' predictions describe the data well: for low modulation frequencies, modulated tones are louder at the same SPL, since the ear can closely follow the slow fluctuations and a value close to the maximum determines the overall loudness perception. For medium modulation frequencies, the ear no longer follows the modulations and a value close to the rms-value determines the overall loudness, which results in level differences close to zero. At high modulation frequencies, the side components in the spectrum of the modulated tone no longer lie in the same auditory filter as the centre component and thus, spectral loudness summation occurs, increasing the loudness of the modulated tone. It can be observed in Figure 3 that the model by Chalupper and Fastl [1] slightly underestimates this effect, which is possibly due to a too an underestimation of spectral loudness summation in this model (see discussion).

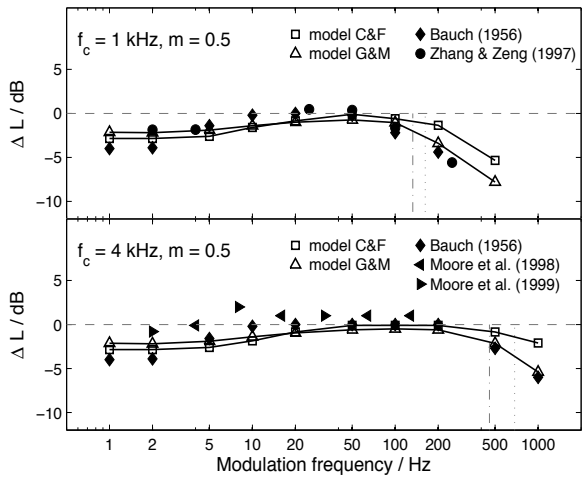

Figure 3: Level difference between amplitude-modulated and equally loud unmodulated tones as a function of modulation frequency for the model by Chalupper and Fastl (open squares) and the one by Glasberg and Moore (open triangles).

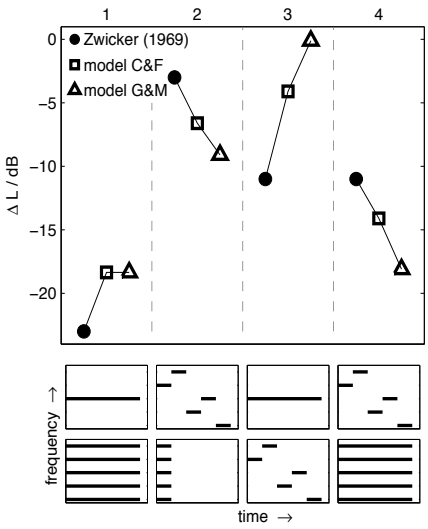

Figure 4: Subset of data by Zwicker [6] (circles). The level difference between equally loud test and reference signals, which are schematically shown in the lower panels, is shown for four conditions along with the model predictions.

Loudness of tone pulses with distinct spectro-temporal patterns

Experiment by Zwicker (1969)

Zwicker performed loudness matches between stimuli of different spectro-temporal patterns [6]. In Figure 4, a subset of four conditions of his data is shown as filled symbols. The level difference between equally loud test and reference signal is indicated for each condition. The lower panels schematically show the spectro-temporal content of the stimuli. The model predictions are represented by open symbols and were derived from the maximum of the short-term loudness. In the first condition, a reference tone pulse of 100-ms duration, a level of 70 dB SPL and a frequency of 1.85 kHz was matched in loudness to a stimulus, which consisted of the sum of five 100-ms tone pulses of frequencies 1000, 1370, 1850, 2500, and 3400 Hz. Each pulse of the latter had the same level and the given level difference was calculated as the difference between the reference level and the level of each of the five pulses. The results indicate that the reference tone had be to considerably higher in level, mainly due to spectral loudness summation. Both models slightly underestimated this effect. The second, third, and fourth condition involved a sequence of tone pulses of 20 ms without inter-pulse pause, in which each pulse had one of the five frequencies mentioned above. This sequence was matched in loudness to a single pulse consisting of all five frequencies (cond. 2), a 100-ms tone at 1.85 kHz (cond. 3), and a 100-ms pulse consisting of all five frequencies. From his results, Zwicker [6] concluded that spectral loudness summation took place even for the sequence of tone pulses with different frequencies, i.e. even when the frequencies were presented non-synchronously. This is only possible, when temporal persistence is assumed in each auditory channel *before* spectral loudness summation takes place. This is supported by the fact that the predictions of the model by Chalupper and Fastl [1], which includes such a mechanism, are closer to the data than those of the model by Glasberg and Moore [2]. However, the effect is not predicted quantitatively.

Own experiment

The effect found by Zwicker [6] was investigated in more detail in the present study by experimentally matching the loudness of a sequence of 10-ms tone pulses of 1.85 kHz to two different test stimuli and several inter-pulse intervals (IPI). In condition 1, the test stimulus was a sequence of five pulses, each with a different frequency. This was similar to the paradigm by Zwicker [6], but here the pulses had the same loudness rather than the same level. In the second condition, each pulse contained all five different frequencies. Figure 5 shows the mean results of 12 normal-hearing subjects plus and minus one standard error (filled symbols), and the corresponding model predictions (open symbols). It can be observed that the level difference in condition 1 disappears relatively quickly in the predictions of both models, while it is always negative in the experimental data even when the individual pulses are separated by 50 ms. In condition 2, the level difference is independent of IPI. The persistence in each channel in the model by Chalupper and Fastl [1] is reflected in the slightly slower decay of the effect in condition 1. However, a comparison to the data shows that the effect is not completely accounted for.

Loudness of environmental sounds

So far, the investigations in this study focused on artificial sounds not usually occurring in environmental conditions. To also compare the model predictions for technical sounds typically encountered in daily life, six sounds were considered, which have the same A-weighted SPL, but vary considerably in their loudness, due to their

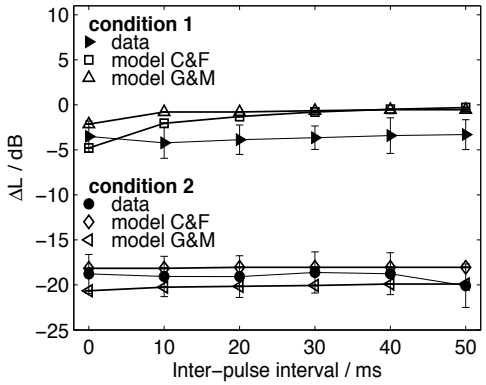

Figure 5: Level difference between equally loud test and reference signals as a function of inter-pulse interval as measured (filled triangles) and simulated by the two models (open symbols).

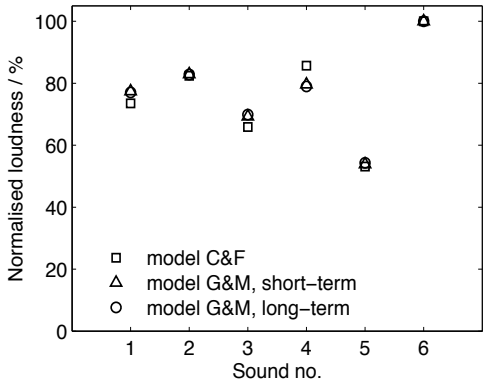

Figure 6: Simulated relative loudness of six environmental sounds with the same A-weighted sound pressure level of 80 dB, normalised to the loudness of sound 6.

different spectral content and temporal structure. The sounds were taken from the website of Prof. H. Fastl [7]. The comparison of the calculated loudness-time functions revealed considerable differences between the two models: the predicted absolute loudness was up to 10 sone larger for some of the sounds in the model by Glasberg and Moore [2] (not shown here). However, as illustrated in Figure 6, the rank correlation between the predicted loudness of the two models is quite high, when the loudness is normalised relative to the loudness of one of the sounds (here sound 6). This is true for both the short-term and the long-term loudness of the model by Glasberg and Moore [2].

Discussion

The present study compared the predictions of two current loudness models, which differ in their dynamic mechanisms, for a set of time-varying sounds. While the model by Chalupper and Fastl [1] and the one by Glasberg and Moore [2] predicted similar results for single tone bursts and amplitude modulated tones, slight differences could be observed for high modulation frequencies. These differences were at least partly due to the reduced spectral loudness summation in the model by Chalupper and Fastl [1]. As mentioned above, this model uses the slightly broader critical-band wide auditory filters, while the model by Glasberg and Moore [2] is based on ERB. Additionally, the compressive exponent in the loudness transformation is slightly larger in the model by Chalupper and Fastl. Both, the broader auditory filters and the reduced compression lead to a smaller effect of spectral loudness summation (see e.g. Verhey and Uhlemann [8]). The differences between the dynamic concepts of the models become apparent when sequences of tone pulses with different frequencies are considered. Zwicker [6] found that spectral loudness summation takes place also for non-synchronous frequency components. This can only be modelled to a very limited degree in the model of Glasberg and Moore [2] due to the finite window lengths used to compute the FFTs. In the model of Chalupper and Fastl [1], this mechanism is explicitly included as persistence in each auditory filter. However, the comparison of its predictions Zwicker's data [6] indicates that the effect is underestimated. This is supported by the results of the present study, which show spectral summation for frequency components separated by up to 50 ms, which cannot be predicted by the model. Even though the dynamic concepts between the two models differ, the rank correlation is very high when technical sounds are considered, even though significant absolute differences can be observed.

References

[1] Chalupper, J. and Fastl, H.: Dynamic Loudness Model (DLM) for normal and hearing-impaired listeners, Acta Acustica united with Acustica, 2002, 88, 378-386

[2] Glasberg, B.R. and Moore, B.C.J.: A model of loudness applicable to time-varying sounds, J. Audio Eng. Soc., 2002, 50(5), 331-341

[3] Zwicker, E.: Über psychologische und methodische Grundlagen der Lautheit (On psychological and methodical principles of loudness), Acustica, 1958, 8, 237-258

[4] Zwicker, E. and Fastl, H.: Psychoacoustics - Facts and models, series in information science, Springer, Berlin, 1999

[5] Moore, B.C.J. and Glasberg, B.R.: A model for the prediction of thresholds, loudness and partial loudness, J. Audio Eng. Soc., 1997, 45(4), 224-239

[6] Zwicker, E.: Der Einfluss der zeitlichen Struktur von Tönen auf die Addition von Teillautheiten (Influence of the temporal structure of tones on the summation of partial loudnesses), Acustica, 1969, 21, 16-25

[7] Fastl,H: http://www.mmk.ei.tum.de/layout.php?selectedMain=Personen&selectedSub=Homes&Special=fas, Technical University of Munich

[8] Verhey, J. L. and Uhlemann, M.: Spectral loudness summation for sequences of short noise bursts J. Acoust. Soc. Am., 2008, 123, 925-934