Modeling roughness perception for sounds with ramped and damped temporal envelopes

Armin Kohlrausch (1,2), Dik Hermes (1), Ronnie Duisters (1,2)
(1) Technische Universiteit Eindhoven, Human Technology Interaction, P.O. Box 513, NL-5600 MB Eindhoven
(2) Philips Research Laboratories, Prof. Holstlaan 4, NL-5656 AA Eindhoven e-mail: {armin.kohlrausch@philips.com}

Roughness is considered to be an important perceptual attribute contributing to overall sound quality. Starting with the work by Vogel, a number of models have been developed to compute the roughness of arbitrary sounds. They consist of a first stage modeling transformations in the peripheral hearing system and a central part in which the envelopes within each auditory filter are analyzed for their specific roughness contributions. These modern implementations, e.g. the one by Daniel and Weber, all have the limitation that they are not sensitive to the phase of the envelope. Specifically, a sound with a nonsymmetrical envelope will receive the same roughness value as its time inverse. Since psychoacoustical data by Pressnitzer and McAdams show a clear difference in perceived roughness for such sounds, we present here extensions to roughness algorithms that are sensitive to the phase of the temporal envelopes. Compared with other implementations, we tested the possible contributions of two specific elements: a compressive gammachirp model for the inner ear, and a variety of hair cell models or black-box stages developed to model adaptation processes. While the gammachirp model in our implementation contributes little to the differences in perceived roughness, all variants of the "hair-cell models" introduced a higher roughness value for sounds with a damped envelope (fast onset, slow decay) than for a ramped envelope (slow onset, fast decay). For the hair cell models the predicted differences were too low, for the black-box model of adaptation they were too high. The implications of such an approach for predicting the roughness for a wider set of perceptual data will be discussed.

1 Introduction

When the envelope of a pure tone is varied sinusoidally, different percepts are observed with increasing modulation frequencies. For very low frequencies the sound will be perceived as a single sinusoid whose amplitude increases and decreases with the modulation frequency. This percept is called fluctuation strength and the periodicity of the sound induces the perception of rhythm. With increasing modulation rate, this percept gradually turns into a percept of roughness. The amplitude fluctuations are no longer perceived separately and the modulations start to sound unpleasant. The unpleasantness has a maximum for modulation rates between 20 and 70 Hz and decreases again for higher rates.

Hermann von Helmholtz [7] considered the relation between roughness and sound quality and used the word 'sensory dissonance', to extend the concept of musical dissonance to non-musical sounds. Later, auditory roughness was systematically studied for stimuli such as amplitude-modulated tones [19], frequency modulated tones [12] and amplitude-modulated noise [5]. A computational model to estimate perceived roughness was developed by Aures [1] and optimized by Daniel and Weber [2].

Our motivation to further develop roughness models comes from the perceptual analysis of sounds created by rolling objects [8]. Figure 1 shows the amplitude of a rolling sound plotted against time. The temporal envelope of this signal is damped as it contains very steep onsets and more slowly decreasing offsets.

![Figure 1: Oscillogram of a rolling sound. The envelope of this signal is damped as it contains very steep onsets and more slowly decreasing offsets.](image)

2 Ramped and damped signals

A prototype of signals with asymmetric envelopes which have been systematically studied in psychophysical ex-
experiments, are so-called ramped and damped signals. Pressnitzer and McAdams [17] used a specific version of those stimuli in a roughness perception experiment. We will therefore adopt their stimulus definition for our model evaluation work. Their stimuli consist of a pure-tone carrier at 2.5, 5 and 10 kHz amplitude modulated by a tone complex $E_{\text{ramp}}(t)$:

$$E_{\text{ramp}}(t) = \sum_{v=1}^{N_{\text{mod}}} \frac{1}{N_{\text{mod}}} \sin(2\pi v f_{\text{mod}} t).$$

(1)

The number of components $N_{\text{mod}}$ was 2, 4 and 7 at the carrier frequencies of 2.5, 5 and 10 kHz. With a modulation frequency of 70 Hz, all components fit within one Equivalent Rectangular Bandwidth around the center frequency: $N_{\text{mod}} \cdot f_{\text{mod}} \leq \frac{1}{2} \text{ERB}(f_c)$. The ramped signal $s_{\text{ramp}}(t)$ can now be constructed by amplitude modulating a carrier signal by the components of $E_{\text{ramp}}(t)$. To obtain a maximum modulation depth of 1 for $m = 1$, the signal $E_{\text{ramp}}(t)$ is normalized by dividing the total signal by its maximum.

$$s_{\text{ramp}}(t) = - \left(1 + m \cdot \frac{E_{\text{ramp}}(t)}{\max(E_{\text{ramp}}(t))}\right) \sin(2\pi f_c t).$$

(2)

A damped signal can be generated by time-reversing the ramped signal:

$$s_{\text{damp}}(t) = s_{\text{ramp}}(-t).$$

(3)

In Fig. 2, a ramped and a damped signal with different center frequencies are shown. It can be seen that a higher center frequency leads to a smoother envelope because the signal consists of more modulation components.

Figure 3 shows the perceptual data of the roughness experiment by Pressnitzer & McAdams. The experiment was performed monaurally with headphones at a level of 60 dB SPL. The ramped and damped stimuli were presented for 1 s each. For each center frequency, all possible combinations of pairs of the data points (ramped/damped, modulation depth) were presented. The listener had to judge which one of the two sounds was rougher. From these paired-comparison data, a roughness scale was computed based on the Bradley-Terry-Luce (BTL) theory (e.g., [4]). Because the experiments for the three center frequencies were carried out separately, the three separate panels cannot be compared quantitatively with each other with respect to the BTL-value. The data show a systematic, monotonic relation between modulation depth and perceived roughness at each center frequency. Furthermore, the damped stimulus is always judged to have a higher roughness than the corresponding ramped stimulus. Although the panels cannot be compared directly due to the applied normalized scale, each single panel provides relative information about the perceived roughness values. For example, at a center frequency of 2.5 kHz the ramped signal with a modulation depth of 0.8 is rougher than the damped signal with a modulation depth of 0.4. At 5 kHz, these two signals have about the same roughness.

### 3 Model components

To model roughness perception, an auditory model was constructed in which four modules can be distinguished. These are represented by the dashed boxes in Fig. 4. The first component is the peripheral filter that approximates the signal processing of the outer and the middle ear. This can be implemented as a bandpass filter. The next two components, the auditory filterbank and the adaptation stage, represent the signal processing in the cochlea. The auditory filterbank is a bank of bandpass filters that divides the wideband input signal into several narrowband auditory channels. The adaptation model can be seen as an envelope extractor with an automatic gain control. Finally, a roughness extraction module generates a roughness value $R$ from the preprocessed signals. As roughness is related to modulation depth $m$ according to the
relation: \( R \propto m^\alpha \), the roughness extraction module estimates the modulation depth of the envelopes in the auditory channels. In the literature, the constant \( \alpha \) takes on values ranging from 1.2 to 2. The total roughness is constructed by the summation of all the partial roughnesses in the auditory channels.

For each module, several implementations presented by different researchers were tested (see Table 1). The following sources were used:


In a first step, we analyzed the contribution of each module to a possible difference in perceived roughness of ramped and damped stimuli. This was done by computing, for all signals used by Pressnitzer and McAdams, the modulation depth at the output of a module. Not surprisingly, the overall bandpass filter simulating the transfer through the outer and the middle ear has no influence. The same is true for the two versions of the inner ear filtering. We should note, however, that our implementation of the compressive gammachirp filter did not include spectral asymmetry and level-dependent filter bandwidth. This bandwidth was, however, controlled by a level parameter which was fixed for a specific signal, so that the filter did not have a dynamically varying bandwidth. A major influence is, however, seen for each of the three modules of hair cell adaptation. They all predict a higher modulation depth at their output for the damped compared to the ramped stimulus, and this difference increases with the modulation depth in the input stimulus.

Figure 5 shows the modulation depth differences \( \Delta m_0 \) for the ramped and damped signals at the output of the three adaptation modules, computed from one period in the middle of a 1-s stimulus. Each panel contains data for a different center frequency and all three signal modulation depths. Overall, the Dau et al. model seems to have the largest effect on asymmetry, while the two hair cell models have the weakest effect. The difference is largest for the signals with a center frequency of 10 kHz. This is probably due to the fact that these signals comprise the largest number of spectral components, resulting in a sharper onset in the acoustic signal and, correspondingly, in a stronger onset response of the adaptation module for the damped signal.

The output of the adaptation module forms the input to the roughness extraction module, for which we tested two implementations. One is based on the work by Aures [1] and the improvements to this model proposed by Daniel and Weber [2]. In this scheme, the envelope is extracted for each auditory channel. While the original implementation used the Bark scale and a triangular filter shape to compute the spread of excitation, our modified implementation was based on the ERB rate scale and adjacent channels had a spacing of 0.5 ERB. The spread of excitation results from the filtering in the initial auditory filter bank. The envelope in each channel is filtered with a modulation transfer function which has a bandpass characteristic with a maximum at about 35 Hz for auditory channels with a low center frequency and at about 80 Hz for high-frequency channels. From the resulting enve-

<table>
<thead>
<tr>
<th>Module</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peripheral filtering</td>
<td>Pflueger et al. [16]</td>
</tr>
<tr>
<td>Auditory filtering</td>
<td>Gammatone [18]</td>
</tr>
<tr>
<td></td>
<td>Compressive gammachirp [10]</td>
</tr>
<tr>
<td>Adaptation</td>
<td>Meddis [14]</td>
</tr>
<tr>
<td></td>
<td>Van Immerseel and Martens [9]</td>
</tr>
<tr>
<td></td>
<td>Dau et al. [3]</td>
</tr>
<tr>
<td>Roughness extraction</td>
<td>Aures [11]</td>
</tr>
<tr>
<td></td>
<td>Synchron. Index Model (SIM) [13]</td>
</tr>
</tbody>
</table>

Table 1: Summary of used modules to form the auditory roughness extraction models. The reference column gives a reference to the literature that describes the implementation of the module in most detail.
lope, the specific roughness is computed per auditory filter, and the total roughness is obtained as a weighted sum across all specific roughnesses.

The other roughness extraction module is based on the synchronization index model proposed by Leman [13]. The synchronization index (e.g. [6]) is a measure that can describe the amount of phase locking in the response of an auditory nerve. The definition used here is obtained by applying the Fourier transform to a firing-pattern record and dividing the value of a certain phase-locked frequency by the DC value. The resulting value is the synchronization index of that frequency. Leman [13] developed two models to estimate roughness by means of the synchronization index. The models make use of the fact that roughness is defined as the energy of relevant modulation frequencies in the auditory channels with respect to the total energy. Both models by Leman make use of the ERB-rate scale containing 40 channels spaced 1 ERB. The model used here evaluates the modulations over all channels before calculating the synchronization index.

Neither of the two roughness extraction modules does contribute to roughness differences for ramped and damped stimuli. The Aures model estimates roughness by calculating modulation depths for each auditory channel, and this is the same for both sounds as long as no preprocessing is applied. The synchronization index model estimates roughness in the frequency domain. The frequency components of the stimuli by Pressnitzer and McAdams are contained within one auditory channel. Since no phase information is processed within a single channel by the synchronization index model, no supplementary effect on the difference in roughness between the two stimuli is expected from this model.

4 Simulation results

Our evaluated module implementations offer twelve possible combinations to form a roughness model, 2 auditory filterbanks, 3 adaptation modules, 2 roughness extraction modules. Besides ramped and damped signals, a large dataset with other stimuli was tested, including sinusoidally-amplitude modulated (SAM) sinusoids, low-frequency pure tones, SAM noise, white noise and frequency-modulated sinusoids for which various parameters were varied. Only a subset of these results will be presented here. All calculations were performed with a sampling frequency \( f_s \) of 48 kHz. First, the roughness values of the reference signals were calculated. If the experiment yielded roughness values on an asper scale, the reference signal is a 1 asper SAM sinusoid which has a center frequency of 1 kHz, a modulation frequency \( f_{mod} \) of 70 Hz, a modulation depth \( m \) of 1, and a level \( L \) of 60 dB SPL.

![Figure 6: Roughness estimate of the ramped and damped stimuli.](image)

![Figure 7: Roughness estimate of the ramped and damped stimuli.](image)

4.1 Ramped and damped sounds

Looking first at the outcome for the ramped and damped stimuli, we remind the reader that only the adaptation stage contributed considerably to differences in predicted roughness. In addition, we observed that the two hair cell implementations also gave similar results, so that it is sufficient to show results for the Dau et al. model (figure 6) and the Meddis hair cell model (figure 7). The asper scales vary for the different plots, but as the BTL scale and the asper scale cannot be quantitatively related, only relative comparisons are legitimate. Comparing the plots in Figs. 6 and 7 with Fig. 3, neither one has a perfect fit. Qualitatively, the influence of modulation depth is predicted correctly in the sense that the roughness of the ramped stimulus is systematically smaller than the roughness of the damped stimulus. But this difference is too large for the Dau et al. model, while the hair-cell models underestimate the roughness difference of ramped and damped signals.

4.2 SAM sinusoids

After having found that all combinations of our modules are able to predict roughness differences between ramped and damped sounds, the roughness of amplitude modulated sinusoids was evaluated for variations in the para-
Roughness of SAM sinusoids. Calculated (closed dots (●) and solid and dashed lines) and perceptually measured roughness (dotted lines) in asper of SAM sinusoids as a function of modulation frequency $f_{\text{mod}}$ and centre frequency $f_c$. $m = 1$, $L = 60$ dB. Modules: gammatone filterbank, Van Immerseel and Martens hair cell model, Aures roughness extraction model. Subjective data reproduced from [2].

Data on the influence of the modulation frequency of the SAM sinusoid are available for several center frequencies. The best fit to the subjective data was obtained with a combination of the gammachirp filterbank, the Van Immerseel and Martens hair cell model and the Aures roughness extraction model. The results are given in Fig. 8. With the Aures roughness extraction model, an unexpected discrepancy occurs for center frequencies of 125 Hz (upper left panel) and 250 Hz (lower right panel). At a modulation rate of zero Hz, the calculated roughness is significantly different from zero. This is due to the fact that the half-wave rectified low-frequency pure tones are treated as envelopes by the hair cell model. The tone frequencies are simply within in the roughness frequency range. The hearing system also appears to have this feature. Recent experiments by Miskiewicz [15] show a significant roughness percept for low-frequency pure tones. This roughness for unmodulated sinusoids is not contained in the experimental data shown in Fig. 8. This discrepancy does not occur for combinations with the synchronisation index model.

The combination of the gammachirp filterbank, Meddis hair cell model and Aures roughness extraction model appears to have the best fit for the relation between roughness and modulation depth. For a modulation rate of 70 Hz, this is shown in Fig. 9. Combinations with the SIM perform worse than combinations with the Aures roughness extraction model, because the former has an incorrect level dependency and the simulation was performed at a level of 70 dB which is 10 dB higher than the level used to calibrate the model to the roughness values of 1 asper.

According to Zwicker and Fastl ([20], p. 260) the roughness of an SAM sinusoid with a modulation depth of 0.98 increases approximately by a factor of 3 when the level is increased from 40 dB to 80 dB. Figure 10 shows the result of the combination of the gammatone filterbank, the Van Immerseel and Martens hair cell model and the Aures roughness extraction model that achieves the best fit for this condition. The combinations with the synchronisation index model, not shown, perform badly as the roughness mostly even decreases with increasing level.

Figure 8: Roughness of SAM sinusoids. Calculated (closed dots (●) and solid and dashed lines) and perceptually measured roughness (dotted lines) in asper of SAM sinusoids as a function of modulation frequency $f_{\text{mod}}$ and centre frequency $f_c$. $m = 1$, $L = 60$ dB. Modules: gammachirp filterbank, Van Immerseel and Martens hair cell model, Aures roughness extraction model. Subjective data reproduced from [2].

Figure 9: Calculated roughness (closed dots (●) and solid line) in asper as a function of modulation depth $m$ for an SAM sinusoid, $f_c = 1$ kHz, $f_{\text{mod}} = 70$ Hz, $L = 70$ dB. The dotted line with the open circles (○) is the power relation $R = 1.36 m^{1.6}$, which is a good approximation for subjective data [2]. Modules: gammachirp filterbank, Meddis hair cell model, Aures roughness extraction model.

Figure 10: Calculated (solid line, closed dots (●) and subjective (open circles (○)) relative roughness of an SAM sinusoid as a function of level $L$, $f_c = 1$ kHz, $f_{\text{mod}} = 70$ Hz, $m = 1$. Modules: gammatone filterbank, Van Immerseel and Martens hair cell model, Aures roughness extraction model. Subjective data reproduced from [20].
5 Conclusions

In this contribution, it was shown how existing roughness models can be extended and modified to be sensitive to the phase of the envelope. A critical component for such an extension is a nonlinear element describing adaptation processes, but none of the three modules examined here resulted in a very good fit between simulations and the experimental data published by Pressnitzer and McAdams. For the hair cell models, the predicted effect was too small, for the black-box model, it was too large. By evaluating the models with a larger data set of perceptual data we found that, overall, the combination of the gammachirp model, one of the two hair cell models and the Aures roughness extraction stage resulted in the best match with the experimental data.

References