Auditory modelling of the perceptual similarity between piano sounds

Citation for published version (APA):

DOI:
10.3813/AAA.919251

Document status and date:
Published: 01/09/2018

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Auditory Modelling of the Perceptual Similarity Between Piano Sounds

Alejandro Osses Vecchi, Armin Kohlrausch
Human-Technology Interaction group, Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, the Netherlands. [a.osses, a.kohlrausch]@tue.nl

Summary
In this study an auditory model which predicts psychoacoustic data was applied to the problem of perceptual similarity between piano sounds. The sounds correspond to loudness-balanced recordings of one note played on seven historical pianos that differ in timbre. The similarity between sound pairs was quantified using a 3-AFC discrimination task. To model perceptual similarity, two expected signals were included in the decision stage of an existing auditory model. The simulations were maximally correlated with experimental data when only the initial part of the sounds (0.2 s) was used as input to the model.

© 2018 The Author(s). Published by S. Hirzel Verlag · EAA. This is an open access article under the terms of the Creative Commons Attribution (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).
PACS no. 43.64.Bt, 43.75.Cd, 43.66.Jh, 43.66.Lj, 43.75.Mn

1. Introduction
In the context of acoustics, similarity assessments are used in sound quality evaluations [1] and in the study of specific sound types [2], among other applications. Since in an everyday listening experience, sound objects are unlikely to be repeated exactly in the same way [3], the concept of similarity is relevant to label sound objects as being the same or not.

In psychology, the similarity between objects is normally assessed experimentally. Often, the goal is to relate the physical properties of the test stimuli to the dimensions of an abstract psychological space [3]. A popular method to construct such a space is the method of triadic comparisons [2, 4]. In a previous study, we proposed an alternative method that assesses the similarity between piano sounds by measuring discrimination thresholds in background noise as signal-to-noise ratios (SNRs) [5]. The focus of the current study is to model such similarity data in noise using an existing computational framework, which has been successfully used to simulate the results of several experimental tasks in noise [6, 7].

2. The auditory model
The family of models that describes the “effective” processing of the auditory system provides a unified framework to simulate a number of auditory phenomena such as masking and modulation-detection tasks [6, 7]. The structure of the adopted auditory model is as described in [7] but with configuration parameters taken from more recent model versions. The block diagram of the model is shown in Figure 1. The peripheral stages of the model (stages 1–6) deliver the internal representation of a sound. The central processor is a back-end stage that compares two or more internal representations with the aim of deciding whether those representations are distinct enough to be judged as different by an artificial listener.

2.1. Peripheral stages
Stage 1. Outer- and middle-ear filtering: This stage is implemented as two cascaded filters whose combined response can be roughly approximated as a bandpass filter centred at \( f_c = 800 \text{ Hz} \) (slopes of 6 dB/octave) (for details see [6]).

Stage 2. Gammatone filter bank: This stage represents an approximation to a critical-band filter bank. The filter bank consists of 30 bands ranging from 101 to 7324 Hz (i.e., 3.4 to 32.4 ERB\(_N\)). The bands are spaced at 1 ERB and one of the bands matches the frequency of 554 Hz (11.4 ERB\(_N\)). The model uses only the real part of the complex-valued all-pole implementation that is described in [8].

Stages 3 and 4. Hair-cell transduction: This stage simulates the transformation from mechanical oscillations in the basilar membrane into receptor potentials in the inner hair cells. The signals are first half-wave rectified and then filtered (5th-order IIR LPF, \( f_{\text{cut-off}} = 770 \text{ Hz} \)). This stage is implemented as in [9].
Stage 5. Adaptation loops: This stage simulates the adaptive properties of the auditory system. Adaptation refers to changes in the gain of the system when the level of the input signal changes. When the input level changes rapidly (relative to five time constants between 5 and 500 ms), the level is transformed linearly. For slower variations, the input level is compressed. This stage is implemented as in [6], but using a stronger overshoot limitation (factor of 5).

Stage 6. Modulation filter bank: In this stage the incoming signal is analysed according to its envelope changes. Each auditory band is split into a maximum of 12 modulation filters with frequencies between 2.5 and 1000 Hz. This stage is implemented as in [6].

2.2. Central processor

In this stage, the information received from the modulation filter bank is compared with a reference representation (template) that is stored in the model. Inspired by the concept of an optimal detector (e.g., part II of [10]), the model can be seen as an artificial listener and the template can be seen as an expected sound representation that gives a clear indication about "what to listen to". The model with the adopted central processor was evaluated for the piano similarity data and its backward compatibility was established for a number of psychophysical intensity-discrimination and masking tasks (shown in [11]).

Template: The use of a template assumes that when discriminating a signal among others, some type of awareness about that signal is expected. This corresponds to a top-down process within the auditory model. This approach is widely used in the field of vision where there is evidence of brain activity in response to features of the expected signal (e.g., [12]). In the model the template is derived (learned) by the artificial listener in the course of a specific experimental paradigm in a condition where the sounds can be easily discriminated (low-noise condition). This is in line with other central processors used in the literature (e.g., [6, 7]). The adopted template derivation is further described in Section 3.1.1.

3. Materials and methods

This study simulates the results of a listening experiment which is similar to the task described in [5]. The same stimuli were adopted in both studies, but reverberation was added to the sounds used in this study.

3.1. Procedure: 3-AFC similarity task

The similarity between sounds was quantified using a 3-AFC discrimination task. Within each experimental trial, one of the sounds served as a reference and was presented in two intervals while the other sound was used as the target and was presented in one interval. The task of the listener was to identify the target interval. A background noise was added to the piano waveforms to change the difficulty of the discrimination task. The level of the noise was adjusted using a two-down one-up rule, which tracks the noise level at which the discrimination score equals 70.7%. The experiment continued until 8 reversals were reached. The starting SNR start was set to 16 dB. The noise level was varied in steps of 4, 2, and 1 dB. The discrimination threshold thresstart was assessed as the median noise level of the last 4 reversals. The presentation level of each interval was varied (roved) between ±4 dB, drawn from a uniform distribution.

3.1.1. Template in the 3-AFC similarity task

In the 3-AFC similarity task, the discriminability of the target sound depends on how different its representation R x is from that of the reference sound R r. For this reason our approach uses two templates: T p,r for the target sound, and T p,r for the reference sound. The templates are obtained as average internal representations at a highly discriminable condition (at SNR supra) and they are normalised to unit energy [7]. We used an SNR supra of 21 dB (i.e., SNR start + 5 dB).

Within a trial, the internal representation of each of the three intervals is cross-correlated with the templates T p,r and T p,r, leading to two decision criteria. Before the actual assessment, the corresponding noise representation R N,x in the interval x is subtracted from the piano-plus-noise representation R x:

\[
\text{CCV}_{x,r} = \frac{1}{T_{x,r}} \int_{t=0}^{T_{x,r}} (R_x - R_{N,x}) \cdot T_{p,r} \, dt, \quad \text{with } x = 1, 2, 3.
\]

\[
\text{CCV}_{x,r} = \frac{1}{T_{x,r}} \int_{t=0}^{T_{x,r}} (R_x - R_{N,x}) \cdot T_{p,r} \, dt. \quad (1)
\]

Two sets of three cross-correlation values (CCV) are used in order to determine the target interval. The artificial listener chooses the interval with the highest CCV x,r and the lowest CCV x,r, to choose the interval with the highest similarity to T p,r and the lowest similarity to T p,r, respectively. A correct decision is obtained if both criteria point to the same interval. The decision is limited by an additive internal noise.
3.2. Stimuli

3.2.1. Piano sounds

Recordings of one note (C#, f₀ = 554 Hz) played on seven Viennese pianos were used. Some information about the piano recordings is shown in Table I. The recordings were auralised using the binaural impulse response of a room having an early decay time (EDTmed) of 3 s. The duration of each sound was set to 2 s, with the note onset occurring at time 0.1 s. The sounds were ramped down using a 300-ms linear ramp. The piano sounds were loudness balanced to have a maximum value of about 18 sone, resulting in sounds with maximum levels between 73.1 and 81 dB SPL. With a set of 7 sounds, the number of possible piano pairs is 21 (see the abscissa of Figure 3a).

3.2.2. Piano-weighted noises

Individual piano noise: To generate noises with spectro-temporal properties similar to those of the piano sounds, an algorithm based on the ICRA noise algorithm [13] was used. As a consequence of a series of interleaved filtering and randomising stages, the algorithm is able to keep the spectro-temporal properties of the input (piano) sounds [5]. This modified algorithm produces noises that are similar to applying a 30-channel noise vocoder to the piano sounds. One realisation of noise N1, generated from piano P1, is shown in Figure 2. The spectrum of N1 is shown as band levels per ERB band. To better visualise the tonal components of P1, its spectrum was obtained using a peak detection algorithm instead.

Paired piano noise: The individual piano noises were not used directly in the 3-AFC task. For a given piano pair, e.g., 13 (piano P1 being compared with P3), one realisation of N1 and one realisation of N3 were combined by averaging the waveforms of the two selected noise realisations. This operation can be seen as a trade-off in the property weighting of, in this example, pianos P1 and P3.

Table I. List of pianos and information about the recording levels as used in this study.

<table>
<thead>
<tr>
<th>ID, Year</th>
<th>Level $L_{\text{max}}$ − $L_{\text{eq}}$ [dB SPL]</th>
<th>Loudness $S_{\text{max}}$ − $S_{\text{eq}}$ [sone]</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1, 1805</td>
<td>80.0 − 67.3</td>
<td>17.3 − 8.8</td>
</tr>
<tr>
<td>P2, 1819</td>
<td>74.4 − 59.2</td>
<td>16.9 − 6.7</td>
</tr>
<tr>
<td>P3, 1828</td>
<td>73.1 − 55.8</td>
<td>17.1 − 6.9</td>
</tr>
<tr>
<td>P4, 1836</td>
<td>78.6 − 64.7</td>
<td>17.1 − 8.6</td>
</tr>
<tr>
<td>P5, 1851</td>
<td>77.5 − 62.9</td>
<td>17.0 − 7.1</td>
</tr>
<tr>
<td>P6, 1851</td>
<td>81.0 − 68.1</td>
<td>18.0 − 8.6</td>
</tr>
<tr>
<td>P7, 1873</td>
<td>80.9 − 69.8</td>
<td>17.4 − 10.1</td>
</tr>
</tbody>
</table>

For each piano pair a set of twelve random (“running”) noises were generated.

3.3. Reference data

The reference data were collected from 20 participants using the stimuli described in this section, providing 210 thresholds. After applying predetermined exclusion criteria, median thresholds $\text{thres}^\text{med}$ were obtained from 6 to 10 data points per pair. The $\text{thres}^\text{med}$ values are shown as red triangles in Figure 3a. The procedure only differed from the description in Section 3.1 in that 12 reversals were used for the staircases.

4. Simulation results

The simulations were run using monaural (left-channel) sounds. Each threshold estimation was repeated 6 times due to the presence of sources of external variability in the 3-AFC task (level roving, running noises) and internal variability of the model (internal noise). The median of the 6 estimated thresholds was used to obtain a single threshold $\text{thres}^\text{sim}$. The initial simulations delivered low SNR thresholds (thres$^\text{sim}$ of −2.75 dB or lower, see Table II, column $t_{\text{obs}}$ = 2 s), meaning that the artificial listener had access to more information than the actual participants. While the model is capable of making point by point comparisons between the reference and target intervals over the 2-s (whole duration) piano sounds, it seems unlikely that humans have access to that same amount of synchronised information. We hypothesised that the simplest way to reduce the available information is by focus-
The need of the model to account for the temporal aspects of the sounds allowed a comparison between two internal representations, two templates were required to allow the artificial listener to distinguish between different pianos. The use of shorter $t_{obs}$ durations is a simple way to reduce the optimally-integrated information in the central processor stage; (2) Although not shown in this paper, the use of shorter $t_{obs}$ brings the interval-CCV values to a range where the source of internal noise limits the performance of the model [11]. It is important to emphasise that the amount of internal noise is not a free parameter of the model but is calibrated in an independent level JND task. The results presented in this paper support the idea that the unified framework of the auditory model can be used to evaluate perceptual tasks using piano sounds. This can be seen as an extension of the use of the model with a central processor that is backward compatible. The choice of model parameters was based on existing literature, and we used only one actual free parameter: the duration $t_{obs}$. An alternative and more elaborate form of information reduction using an additional source of internal noise is described in [14]. Such an approach is compatible with the adopted auditory model and might have replaced the use of $t_{obs}$. The rationale is, however, the same in both approaches.

Acknowledgements

We would like to thank Kevin Tang for collecting the experimental data presented in this study. This study was funded by the European Commission within the ITN Marie Curie Action project BATWOMAN under the 7th Framework Programme (EC grant 605867).

References


