Prediction of perceptual similarity based on time-domain models of auditory perception

Alejandro Osses

PREDICTION OF PERCEPTUAL SIMILARITY BASED ON TIME-DOMAIN MODELS OF AUDITORY PERCEPTION

ALEJANDRO OSSES VECCHI

Prediction of perceptual similarity based on time-domain models of auditory perception

Alejandro Osses Vecchi

The work in this dissertation was financially supported by the European Commission within the ITN Marie Skłodowska-Curie Action project BATWOMAN under the 7$^{\text{th}}$ Framework Programme (EC grant agreement Nr. 605867).

A catalogue record is available from the Eindhoven University of Technology Library.

ISBN: 978-90-386-4550-6

NUR: 776

Keywords: Perceptual similarity, auditory modelling, musical acoustics

Cover design: Carolina Osses Vecchi

Printed by: ProefschriftMaken ‖ www.proefschriftmaken.nl.

Prediction of perceptual similarity based on time-domain models of auditory perception

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op woensdag 19 september 2018 om 11.00 uur

door

Alejandro Alberto Osses Vecchi

geboren te Santiago, Chili

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. G.J.J.A.N. van Houtum
1^{e} promotor:	prof.dr. A.G. Kohlrausch
2^{e} promotor:	prof.dr. A. Chaigne (ENSTA ParisTech)
leden:	prof.dr. T. Dau (Danmarks Tekniske Universitet)
	prof.dr.-ing. M. Kob (Hochschule für Musik Detmold)
	dr.ir. R.H. Cuijpers
	dr.ir. M.C.J. Hornikx

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Summary

Title: Prediction of perceptual similarity based on time-domain models of auditory perception

Objects or situations in an everyday context are unlikely to be experienced twice in the same way. The more exposed an individual is to a given object or situation, the more familiar he or she becomes with that object or situation. While listening to a sound object, we may find that it resembles another sound with which we are familiar. In this case we may label both sounds as being "similar". Similarity assessments may indicate whether two or more sound stimuli share common perceptual properties. Let us consider a sound quality evaluation between the reference sound A and the test sound B. The test sound B can be chosen as being (1) a modified version of A, (2) a synthesised version of A, or (3) a sound that is believed to be similar to A. An evaluation of the first type (1) is useful to study which properties of sound A are perceptually prominent. An evaluation of the second type (2) can be used to validate a computational model that accounts for the theory that is believed to be relevant to recreate sound A. An evaluation of the third type (3) can lead to a measure of perceptual distance between sounds A and B. The work in this dissertation is mainly concerned with this latter type of evaluation.

The goal of this research work was to gain insights into human performance in a similarity task. For this purpose, the similarity of a set of sounds was first experimentally assessed. Subsequently, the same experimental framework was implemented and used as input to a state-of-the-art model of auditory perception. The hypothesis was that the similarity assessments obtained from the auditory model are significantly correlated with those obtained experimentally.

In this study we chose to compare sounds using the internal (sound) representations delivered by an auditory model. The model, referred to as perception model (PEMO), offers a unified framework that has been successfully used to simulate a number of auditory phenomena such as masking and modulation tasks. The advantage of using a unified framework is implicitly emphasised in Chapter 2, where recorded and synthesised sounds of an instrument called Hummer are compared (type 2 task) using three auditory models that deliver four psychoacoustic descriptors: Loudness, loudness fluctuations, fluctuation strength,

and roughness. The model estimates are compared using the concept of just-noticeable difference (JND), with one JND value for each of the four psychoacoustic descriptors. If the descriptors differ by less than one JND, the sounds are considered to be perceptually identical along the evaluated dimensions.

In Chapter 3 a new method to assess the perceptual similarity between sounds is introduced and validated. In the so-called instrument-in-noise method two sounds are compared using a three-alternative forced-choice paradigm (3-AFC). The reference sound is presented twice and the test sound is presented once. The task of the participant is to identify in which of the three sound intervals the test sound was played. One of the key aspects of this method is that a background noise is added to manipulate the difficulty of the task. This allows to assess the similarity between two sounds as a performance task. The background noise needs to have similar spectro-temporal properties to those of the test sounds. For this purpose a noise generation algorithm similar to the ICRA noises was adopted. Two sounds that are similar tolerate a low background (ICRA) noise to correctly discriminate one from the other in contrast to the case of two sounds that are more dissimilar, where more (ICRA) noise needs to be added before the participant's performance decreases. The sound stimuli consisted of recordings of a single note from seven historical pianos. With seven sound stimuli, 21 possible piano pairs can be evaluated. Twenty participants were asked to compare those 21 piano pairs using two methods: (1) the instrument-in-noise method, and (2) the method of triadic comparisons. The discrimination thresholds from the instrument-in-noise method were significantly correlated with the similarity assessment obtained from the method of triadic comparisons.

In Chapter 4 the participant's performance for the instrument-in-noise test is simulated using the same piano sounds and experimental paradigm as in Chapter 3 but using an "artificial listener". The artificial listener uses internal representations obtained with the PEMO model and decides whether two representations are distinct enough to be judged as "different". This decision is based on the concept of optimal detector taken from signal detection theory. Both, the peripheral stages (that deliver the internal representations) as well as the central stage (the artificial listener) of the PEMO model are described in detail in this chapter. The discrimination thresholds obtained with the PEMO model are significantly correlated with the experimental thresholds.

In Chapter 5, the same seven piano sounds of Chapters 3 and 4 but considering a reverberant environment (early decay time of 3.0 s) were perceptually evaluated. Discrimination thresholds obtained from twenty new participants were assessed and subsequently simulated using the PEMO model. The results had a similar (significant) correlation between experimental and simulated thresholds, as observed when comparing the results of Chapters 3 and 4.

In Chapter 6 a binaural model that has the same peripheral stages as the PEMO model, but using a different central processor, is used to simulate the perceived reverberation (reverberance) of orchestra sounds in eight different acoustic environments. The main goal of this chapter is to show one example of application that further extends the use of the auditory models. The reverberance estimates obtained from the binaural model were compared with the experimental results of a multi-stimulus comparison task. The experiment considered 8 instruments and they were evaluated by 24 participants. The multi-stimulus comparison is an alternative and faster way to compare sounds pairwise and it can be used to develop perceptual scales. The experimental reverberance estimates were significantly correlated with the simulated reverberance estimates.

The work presented in this dissertation supports the use of a unified auditory modelling framework to simulate a perceptual similarity task using sounds that are non-artificial. The unified framework was used to evaluate two similar sets of sounds: single-note recordings from seven piano sounds without (Chapters 3 and 4) and with reverberation (Chapter 5). The experimental paradigm, that we named instrument-in-noise test, can be further used to evaluate other musical instruments as far as the sounds to be evaluated have the same duration and are tuned to the same frequency. These aspects are relevant to appropriately generate noises that match the spectro-temporal properties of the sounds being tested.

Table of contents

List of acronyms and abbreviations

AFC	Alternative forced-choice
AM	Amplitude modulation
AMT	Auditory Modelling Toolbox
BBN	Broadband noise
BPF	Band-pass filter
BRIR	Binaural room impulse response
CCV	Cross-correlation value
dBFS	dB Full scale
DLM	Dynamic loudness model
DR	Dynamic range
EDT	Early decay time
ERB	Equivalent rectangular bandwidth
F0	Fundamental frequency
FFT	Fast Fourier transform
FIR	Finite impulse response
FM	Frequency modulation
FS	Fluctuation strength
HPF	High-pass filter
ICRA	International Collegium of Rehabilitative Audiology

IFFT	Inverse fast Fourier transform
ISO	International Organisation for Standardisation
IIR	Infinite impulse response
IQR	Interquartile range
JND	Just-noticeable difference
LPF	Low-pass filter
MDS	Multidimensional scaling
MU	Model Units
PEMO	Perception model
R	Roughness
RC	Resistor-Capacitor
RAA	Room Acoustic Analyser
RMS	Root mean square
RT	Reverberation time
SNR	Signal-to-noise ratio
SPL	Sound pressure level
STFT	Short-time Fourier transform
TVL	Time-varying loudness

1 General introduction

1.1 Sounds as internal representations in the auditory system

The sense of hearing provides us with the possibility to explore and interact with our surrounding sound environment. Examples of this interaction are the ability to localise a sound object or to obtain information about its identity. The ability to access such information by using our hearing system is hypothesised to be possible due to the existence of internal processes of perceptual organisation (McAdams & Bigand, 1993). The information used by these internal processes is what we call "internal representation". Internal representations are sometimes referred to in the literature as "mental representations". This term indicates that the auditory system delivers information about the sound object to the brain. The hearing system consists of a "mechanical" part –comprising the outer, middle, and inner ear– and a "neural" part. After the mechanical or peripheral auditory processing the sounds are represented as firing patterns in the auditory nerve. The neural part comprises the connectivity and involved functional mechanisms that transmit the information, i.e., firing patterns of the auditory nerve, through the central nervous system to the brain (see, e.g., Kohlrausch et al., 2013).

There is consensus that the neural activity in the auditory nerve is encoded according to a frequency-to-position conversion that occurs in the inner ear (see, e.g., Greenwood, 1990; Robles & Ruggero, 2001). This frequency-position mapping is known as the tonotopic organisation of the cochlea. The mechanical part of the auditory system is therefore simulated as a set of band-pass filters. In the study by Saremi et al. (2016) seven of such filter banks have been reviewed and compared in terms of their capability to reproduce relevant aspects of the cochlea.

Table 1.1: Selected list of central processors (sorted by year of publication) that are used as back-end stage for published computational models of the auditory periphery. The column "Nr. of Repr." indicates the number of representations required by the "criterion" of the central processor.

Central processor type	Nr. of Repr.	Peripheral stage based on
A. Optimal detector (Dau et al., 1997a)	3	Dau et al. (1997a)
B. Autocorrelator-based pitch analyser (Meddis & O'Mard, 1997)	1	Meddis and Hewitt (1991)
C. Discriminability analyser (Fritz et al., 2007)	2	Glasberg and Moore (2002)
D. Envelope analyser (Jørgensen & Dau, 2011)	1*	Ewert and Dau (2000)
E. Room Acoustic Analyser (van Dorp, 2011)	1	Breebaart et al. (2001)
F. Envelope analyser (Mao & Carney, 2015)	1	Zilany et al. (2009)

(*)Processor D processes "individual" speech samples in noise (i.e., one test interval), but the processor also needs to have access to the internal representation of the noise alone in order to generate its output metric.

In contrast to the processing in the peripheral auditory system, there is no similar consensus with respect to stages of higher-level neural processing. This has generated diverging approaches to further process the firing patterns of the auditory nerve and, therefore, to obtain and use internal representations.

Computational models of auditory processing normally consist of the stages of peripheral and central processing. The peripheral processing stage represents the mechanical part and initial stages of neural processing of the auditory system. The central processing stage is used as a back-end module for the peripheral processing. A selected list of central processors attached to published models of the auditory periphery are presented in Table 1.1. A central processor accounts for: (1) high-level neural processing of the auditory system (to a greater or to a lesser extent), and (2) coupling of the internal representation to a certain "criterion" (decision stage) that provides concrete information about the processed sound object. In general this latter aspect is assessed by either comparing two or more internal representations (see, e.g., processors A and C in Table 1.1) or by converting the internal representation into a metric believed to reflect some perceptual aspect of the processed sound object (see, e.g., processors B, D, E, and F in Table 1.1). In this dissertation a computational model that follows the former rationale is used. We use an updated version of the model described by Dau et al. (1997a) with a central processor that compares different internal representations by using the concept of optimal detector (see Chapters

4 and 5)[1]. Therefore, our work is concerned with one possible way of comparing internal representations of different sounds. Particularly, the comparison of internal representations is implemented as a performance task and it is applied to the evaluation of perceptual similarity between complex sounds.

As test stimuli, musical instrument sounds are used. This choice is motivated by: (1) the complex nature of the sounds, (2) the fact that musical instrument sounds have been thoroughly studied in physical acoustics, and (3) the fact that the auditory model used in this thesis has been primarily applied to study artificial sounds (see, e.g., Dau et al., 1996a, 1996b; Jepsen et al., 2008) and speech (see, e.g., Holube & Kollmeier, 1996; Hansen & Kollmeier, 2000; Jørgensen & Dau, 2011) and less often to other types of sounds, including musical instrument sounds (Huber & Kollmeier, 2006). Although Huber and Kollmeier applied the auditory model to more diverse sets of sounds, their central processor was adapted to provide a quality metric and, therefore, the goal in their study was to assess judgements of sound quality rather than simulating performance. In this context, the work presented in this thesis can be seen as a possibility to extend the use of the unified framework offered by the auditory model.

In the next section, a definition of what we understand as sound complexity is given. This is followed by a review of the experimental procedures used to perceptually compare sounds. A special emphasis is given to methods that use a discrimination threshold approach. This is because the simulations of perceptual similarity that are to be presented in Chapters 4 and 5 are based on a similar rationale to that of previous simulations using a discrimination threshold approach.

1.2 Musical instruments as complex sounds

According to Yost et al. (1989), three of the properties that characterise the perception of complex sounds are: (1) Spectral complexity, (2) temporal complexity, and (3) noise embedment. The spectral complexity refers to the presence of more than one frequency component in a sound. The temporal complexity indicates that the spectral as well as the temporal characteristics vary over the duration of the sound. Finally, the target sound object is embedded in an acoustic environment consisting

[1] As an extension to the same modelling scheme, an example of a central processor that transforms the internal representations into a metric of reverberation, which is based on central processor E (see Table 1.1), is given in Chapter 6.

of more objects. The "other objects" constitute a background noise that affects directly or indirectly the sound object properties.

According to these definitions, the sets of sounds used throughout this dissertation are both spectrally and temporally complex. Since all the stimuli correspond to recorded musical instruments and they are noise-free, the role of noise embedment will not be addressed here. Noise embedment will be used, however, to mask the properties of given target sounds. Those noises are of stochastic nature, but have the same spectro-temporal characteristics as the target sounds. The generation of such noises is described in Chapters 3 and 5.

A spectro-temporal representation of three sounds is shown in Figure 1.1. The sounds correspond to a 1000-Hz pure tone (panel A), a recording of an instrument called Hummer, resonating in its acoustic mode 2 (panel B), and a recording of a piano sound, note C#$_5$ (panel C). The Hummer corresponds to the test instrument studied in Chapter 2 and the piano (note C#$_5$) corresponds to the test instrument studied in Chapters 3, 4, and 5. In the top panels of the figure the respective waveforms (black lines) together with their Hilbert envelope (red lines) are shown. The envelope is used as a representation of the slow response of the human hearing system to incoming sounds. This characteristic is sometimes referred to as "sluggishness" of the hearing system. Therefore, a constant envelope can be interpreted as belonging to a steady sound. Likewise, an envelope that varies in time is attributed to a sound that is perceived as a time-varying waveform. In the bottom panels of Figure 1.1 a short-time Fourier transform (STFT)[2] analysis is shown. Darker regions in the spectrogram represent higher signal amplitudes. Those amplitudes range between the maximum in the signal (darkest area) down to a floor amplitude that is 50 dB below (white area). The red lines indicate the estimated fundamental frequency (F0) of the signals. The frequency range in each panel was chosen to facilitate the visualisation of the relevant spectral components in the sounds.

According to our definition of complexity, the sounds in panels A, B, and C of Figure 1.1 have an increasing complexity. The sine tone consists of a single spectral component at a frequency of 1000 Hz and its envelope is steady. The hummer sound has an F0 of 430 Hz, with a

[2]For the STFT analysis the waveforms were downsampled to an f_s of 22050 Hz. The STFT is based on successive 32768-point FFTs performed on 40-ms signal segments (zero-padding was applied) with 75% overlap (10-ms hop size). The resulting frequency resolution of the analysis is 0.7 Hz.

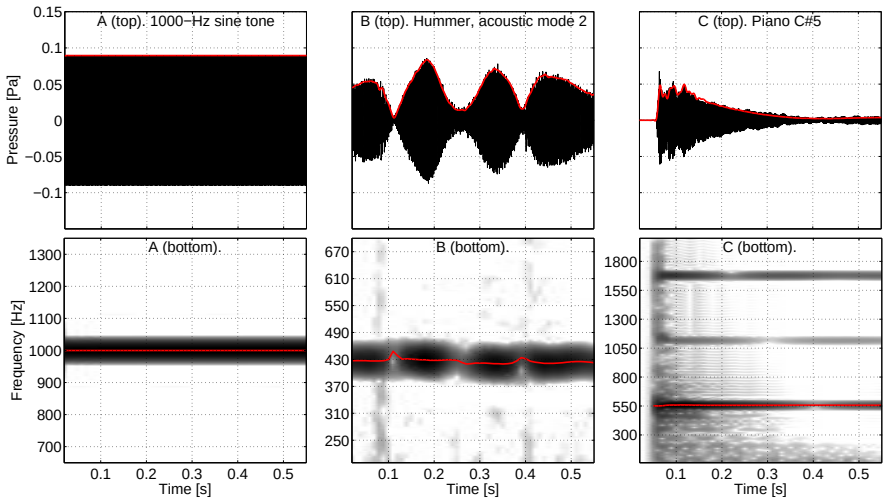


Figure 1.1: Spectro-temporal analysis for three different sounds: (Panel A) A 1000-Hz pure tone, (Panel B) a Hummer sound, and (Panel C) a piano sound. In the top panels the sound waveforms are shown together with their Hilbert envelope (red lines). In the bottom panels, an STFT analysis is shown (Δf=0.7 Hz, 40-ms analysis frame, 10-ms hop-size). Dark regions indicate higher signal amplitudes, the dynamic range corresponds to 50 dB. The red lines indicate the F0 of the sounds. The F0s of the pure tone and piano sound are 1000 Hz and 554 Hz, respectively. The F0 of the hummer sound varies between 419 and 448 Hz.

frequency variation between 419 and 448 Hz and it has a time-varying envelope with amplitudes between 48.9 dB ($p = 5.6$ mPa at $t = 0.11$ s) and 72.5 dB ($p = 84.3$ mPa at $t = 0.18$ s). The piano sound has more complex spectro-temporal characteristics. In terms of frequency (panel C, bottom), the F0 of 554 Hz, the first two partials (around $f_1 = 1110$ Hz and $f_2 = 1660$ Hz) and several (less strong) frequency components are visible in the figure. The less-strong broadband frequency components correspond to the so-called attack noise and they decrease rapidly in amplitude after the note onset. Higher frequencies vanish more quickly in comparison to the lower frequencies. As can be seen in panel C (top), the signal has a strong onset with an amplitude that increases up to 70.5 dB ($p = 67.3$ mPa at 0.07 s) within 10 ms.

For the interested reader, the (complex) spectro-temporal characteristics of 25 musical instruments can be found in Chapter 3 of the book by Meyer (2009). In that review, selected notes of each instrument and their development in time in a three-dimensional pattern (time-frequency-amplitude) are shown. For the particular case of the piano, a C_6 note (F0 of 1047 Hz) is described in detail. The analysis also includes a descrip-

tion of how the intensity and the style of playing (*legato* and *staccato*, for note C_3) affects the tone colour of the resulting sound. These aspects may also be applicable (but they are not discussed in this thesis) to our test piano recordings (note $C\#_5$), especially the description of the attack noise given for C_6 due to its proximity to the $C\#_5$ string (less than one octave difference).

1.3 Methods for the perceptual evaluation of musical sounds

In this section, we review the most relevant approaches used so far to evaluate aspects of sound perception applied to musical sounds. A more detailed description is provided for those methods that have been directly or indirectly used in this dissertation. Other comprehensive reviews of experimental methods used in psychophysics are given by McAdams and Bigand (1993, Chapter 6) and by Kingdom and Prins (2016).

In line with the review given by McAdams and Bigand in the context of classification and recognition of sound sources, the different experimental tasks can be grouped in one of the following types: (1) Discrimination, (2) Psychophysical rating scales, (3) Preference/similarity ratings, (4) Matching, (5) Classification, and (6) Identification. For each of these tasks one or more experimental methods can be used. Based on the expected outcome of each method, the described tasks are either labelled as a "performance" or as an "appearance" method. This label responds to whether the trial responses can be evaluated as "correct/incorrect" or not. In an appearance-based method, apparent magnitudes (that are relative or absolute) along any specific dimension or stimulus attribute are collected.

1.3.1 Discrimination

Category: Performance – threshold methods

In this task the participant is asked to differentiate between two or more stimuli. The percentage of correct responses is calculated for different levels of the independent variable. The task can be implemented as an m-alternative forced-choice (AFC) experiment. In an m-AFC experiment there are m intervals per trial and m alternatives from which the participant has to choose one. In a 2-AFC task, the participant needs an explicit reference to the dimension being investigated and he/she has to be somehow familiar with it. For instance, in a 2-AFC intensity discrimination task, the participant is asked: "Which of the two intervals does

sound more intense?" (see, e.g., Rabinowitz, 1970, intensity discrimination with pure tones). In this case, it is expected that the participant is familiar with the concept of intensity. The implementation of the task as a 3-AFC experiment opens the possibility to not explicitly ask the participant about the dimension being investigated. In the example of intensity discrimination, the question may turn into "Which of the three intervals does sound different?".

1.3.2 Psychophysical rating scales

Category: Appearance – scaling methods

In this task the participant is asked to ascribe a number to the sensation produced by a given stimulus. The goal is to construct an interval scale related to a specific sensation along which the set of stimuli can be ordered from low to high. The method of magnitude estimation provides one way to construct such a scale. This method has been used mostly to develop scales of basic auditory sensations such as loudness (Stevens, 1955, 1956; Houtsma et al., 1987), fluctuation strength (Fastl, 1982, 1983; García, 2015), and roughness (Fastl, 1977; Kemp, 1982). Three (existing) psychoacoustic models that have been developed based on the scales of loudness (Chalupper & Fastl, 2002), fluctuation strength (García, 2015; Osses et al., 2016), and roughness (Daniel & Weber, 1997) are used in Chapter 2 to evaluate a musical instrument called hummer.

1.3.3 Preference/similarity ratings

Category: Appearance – forced-choice scaling methods

Pairwise and triadic comparisons

In this type of tasks the participant is forced to make a choice out of a given number of m stimuli. When comparing the stimuli pairwise ($m = 2$), one possible task is to indicate the preference between two stimuli. In this case there is no explicit reference about the dimension being investigated. In a triadic comparison ($m = 3$) the participant is asked to indicate the pair of sounds that may be grouped together when being compared. Therefore, the only instruction is to base their choice on how similar the stimuli within a trial are. The participant's choices are collected into a matrix, that is referred to as preference (if $m = 2$) or similarity matrix (if $m = 3$). A processing of the scores in the matrix should result in an interval scale. One of the methods used to generate such a scale is the so-called multidimensional scaling (MDS) (Kruskal, 1964a, 1964b). The MDS method provides a way to visualise the distribution

of the test stimuli in a multidimensional (abstract) space. The interval similarity scale is derived by assessing the distance between pairs of stimuli in the resulting space. In the context of auditory perception, triadic comparisons have been used to evaluate artificial complex tones (Levelt et al., 1966), the similarity between music genres (Novello et al., 2011), and the similarity of violins with different vibrato amplitudes (Fritz et al., 2010). Pairwise comparisons have also been used in the evaluation of musical instrument tones (Grey, 1977; Grey & Gordon, 1978) and timbre variation in monophonic and polyphonic contexts (Grey, 1978).

Multi-stimulus comparison

The method of multi-stimulus comparison (De Man & Reiss, 2013) is an alternative to pairwise comparisons. In this task, the participant is asked to distribute multiple sound stimuli along a single scale. In this way, multiple stimuli are evaluated within one trial. The multi-stimulus comparison is very similar to the "Multi-stimulus test with hidden reference and anchor" (MUSHRA) (ITU-R, 2015), but it does not require the use of a reference nor (necessarily) anchors. An example of a multi-stimulus comparison is given in Chapter 6.

1.3.4 Classification

Category: Appearance – scaling methods

In this task the participant is asked to group the stimuli based on "a criterion". The criterion is often freely defined by the participant. As result, each category is defined by a freely-defined label and the stimuli are distributed along this label scale. For this reason, the task is also known as free categorisation. A free categorisation task can be interpreted as a way to obtain an individualised scale, because the label can vary from participant to participant. In general, the classification requires more than one label (leading eventually to more than one scale). Since the labels (i.e., the judgement criteria) are defined by the participants, the interpretation of the resulting scale is facilitated. An example of free categorisation is given in the perceptual evaluation of violins by Saitis, Fritz, Scavone, Guastavino, and Dubois (2017). In their study, 30 experienced violin players were asked to rank either 8 or 10 violins providing written responses to justify their choices. The analysis of the written responses lead to 828 words linked to concepts of violin quality. A subsequent analysis of semantic proximity allowed to group the words into 8 semantic categories, which the authors linked to timbre, intensity, and playability characteristics of the violins. The concept of "category"

is comparable to the concept of "dimension" of a perceptual space (that can be obtained with MDS) with the difference that the latter one is of an abstract nature and requires further interpretation.

1.3.5 Identification

Category: Performance

In an identification or recognition task the participant is asked to link the test stimuli with names or labels. The identification task can be based on open-set labels (free identification) or on close-set labels. Possible analyses for an identification task are: (1) the assessment of identification scores (see, e.g., Saldanha & Corso, 1964), (2) the construction of confusion matrices (see, e.g., Steeneken, 1992, his Chapter 3), and (3) the measurement of reaction times (see, e.g., Agus et al., 2012).

In the study by Saldanha and Corso, notes of 10 musical instruments were recorded and presented in their original form and with 5 different types of modification. The participants had to identify the instrument being played based on a closed set of labels. Although the authors were able to draw conclusions about the instruments that were easier to identify and the type of modification that lead to a better performance, overall low scores per instrument were obtained (only three instruments had identification scores above 50%). The authors argued that a more elaborate analysis of the incorrect scores would have provided further information to better explain their results. They indicated, for instance, that in most of the wrong answers for violin sounds, the cello had been chosen and that this information could not be observed by only using identification scores. An analysis that can reflect this information is the construction of a confusion matrix. Such a matrix is constructed by counting the number of times each stimulus is chosen over the other. A high confusion score provides evidence of shared (perceptual) stimulus features. In this way similarity can be implicitly evaluated. This gives the possibility to analyse confusion matrices using techniques as principal component component analysis (PCA) and MDS.

1.4 Linking methods of perceptual evaluation with auditory modelling frameworks

Our interest in this thesis is, as pointed out in Section 1.1, to evaluate the similarity between sounds by comparing their internal representations which, in turn, are derived from an auditory model (Dau et al., 1997a).

The decision stage of the model compares the internal representations in terms of their spectro-temporal distribution of neural activity, which is obtained from the corresponding sound intervals, usually presented in 3-AFC trials.

In order to implement a similarity task using the same 3-AFC paradigm, the question to the participant needs to be implicitly asked. One way to do this would be to implement the experimental procedure as a **discrimination task** ("which of the three sounds is different from the other two?"). Considering the definitions of the previous section, such a task corresponds to a performance task with forced choices. Other methods that may be applicable to implement our similarity task are: the method of triadic comparisons, and an identification task. The reasons to favour the implementation of the similarity experiment as a discrimination task over those methods are:

- The **triadic comparison** method is an appearance task, i.e., there are "no wrong answers" in the similarity judgement;
- The similarity (distance) measure in the triadic comparisons depends on the choice of the set of stimuli, and;
- Although the participant's performance can be assessed in an **identification task**, this performance may also be influenced by the set of stimuli (or stimulus labels) chosen for the experiment.

Judgements of similarity in a 3-AFC discrimination task would only depend on the two sounds being compared (presented in three intervals) and will not be influenced by the "other" sound stimuli of the dataset. Additionally, the performance can be quantified by the percentage of correct responses (scores), and the question "which of the three sounds is different from the other two?" can be evaluated by the auditory model in terms of the spectro-temporal characteristics of each sound interval, under the assumption that similar sounds have similar spectro-temporal characteristics. If the discrimination task is implemented using an adaptive procedure, the independent variable (the adjustable parameter) is chosen to influence the difficulty of the task, and discriminability thresholds can be obtained. An example of such an approach is the study on violin sounds by Fritz et al. (2007), where the independent variable was a gain applied to the test sound in four different frequency regions. This lead to the estimation of four amplitude thresholds. They used an auditory model –the multichannel excitation-pattern model (Moore & Sek,

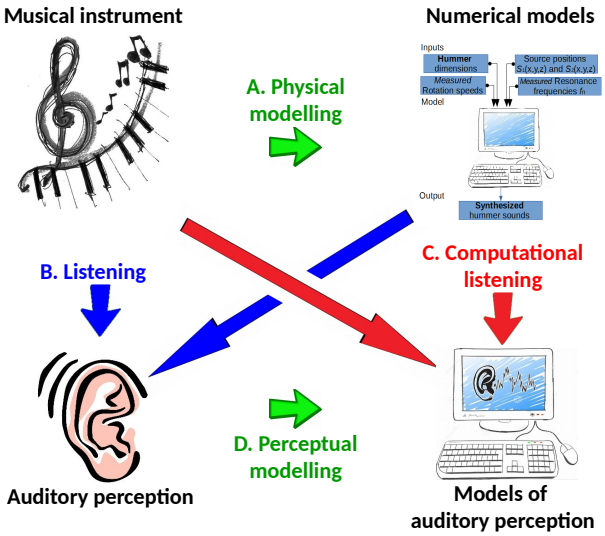


Figure 1.2: Schematic drawing of possible steps to study the properties of a sound source. In this particular example the sound source is a musical instrument.

1992; Glasberg & Moore, 2002) (Processor C in Table 1.1)– to simulate the amplitude thresholds of five of their participants. They succeeded to recreate the experimental thresholds for two test notes (G_3 and E_5), with a deviation of less than 1 dB. These results served to evaluate which of three possible ways of combining information across auditory frequency channels was adopted by their participants.

We adopt a similar approach to that used by Fritz et al. (2007). Our auditory task is implemented as a discrimination experiment, and its results are compared with simulated thresholds using an auditory model with the **goal** of understanding what type of auditory information do participants use when comparing our test (piano) sounds. The independent variable in our approach is a carefully chosen background noise rather than the use of a direct modification of the (piano) waveforms.

1.5 Motivation of this thesis

When studying a musical instrument, possible approaches to investigate its properties can be summarised using the diagram of Figure 1.2. The approaches are classified into one of the following types: (1) Physical modelling, (2) listening, (3) computational listening, or (4) Perceptual modelling. In Section 1.3, a review of methods adopted in the "listening" approach has been given. Although this has not been pointed out

so far, due to the (on average) long time required to conduct listening experiments, an alternative is to use the approach that we labelled as "computational listening", which represents the use of acoustic or psychoacoustic metrics obtained from dedicated computer programs. A very simple example of computational listening is the comparison of two STFTs. A more elaborate example is given by the acoustic similarity metric of Agus et al. (2012), which is based on an energy average using a simplified internal representation of the sounds (Moore, 2003). The authors used this information to explain the results of their identification test, where shorter reaction times were found when the task considered less similar sounds.

The "physical modelling" approach relies on the simulation of a sound source by implementing a model for its vibration and sound radiation. Two examples of this approach in the study of guitar and piano sounds are given by Derveaux, Chaigne, Joly, and Becache (2003) and Chabassier, Chaigne, and Joly (2013). In order to evaluate how well does a given numerical model match –or how similar the simulated sounds are to– the sound source under evaluation, a comparison with actual recordings should be conducted. The comparison can be done by either running listening experiments ("listening") or by applying some kind of computer analysis ("computational listening").

The remaining part of the diagram, i.e., the "perceptual modelling" approach, constitutes the **main goal** of this thesis. This approach consists of gaining insights into human performance –in our case, into "how discriminable" two sounds are– by incorporating advanced perceptual aspects into a computational listening approach. We compare experimental thresholds with simulated (or "perceptually modelled") thresholds obtained from an auditory model. The test sounds in our task are individual piano notes (Chapters 3 to 5). As an "acoustic event", individual notes are considered to be one of the simplest cases to study (McAdams & Bigand, 1993) when compared with the use of melodic lines or a fragment of music with multiple instruments. Our efforts are focused, however, on the complex nature of the piano sounds and on a detailed analysis of their (multidimensional) internal representations obtained from an auditory model. This model corresponds to an updated version of the perception model (PEMO) described by Dau et al. (1997a). As a consequence of using the PEMO model to assess simulated thresholds for complex (piano) sounds, the work in this thesis can be seen as a further extension of this unified modelling framework that

has already been successful in simulating human performance in a range of auditory tasks.

1.6 Outline

In Chapter 2 a selection of psychoacoustic descriptors is reviewed and applied to a set of sounds. The descriptors correspond to the classic psychoacoustic measures of loudness, roughness and fluctuation strength. The descriptors are used to compare sounds of a musical instrument called hummer. The hummer is a plastic corrugated pipe that generates sounds when being rotated at specific speeds. In this chapter existing recordings of the hummer (Hirschberg et al., 2013) are quantitatively compared with a computational model of the hummer (Nakiboğlu et al., 2012). This study case corresponds to an example of the "computational listening" approach shown in the schema of Figure 1.2, with as result an evaluation of the numerical model of the instrument.

In Chapter 3 an experimental method to assess the perceptual similarity among sounds is presented. The experimental method corresponds to an "instrument"-in-noise discrimination test where the noise is used to manipulate the difficulty of the discrimination. The method of triadic comparisons –largely used in psychology– is used as reference method. A perceptual similarity study using recorded piano sounds of one note played on a number of historical pianos is presented. The instrument-in-noise method provides discrimination thresholds, expressed as signal-to-noise ratio (SNR), that are significantly correlated with the Euclidean distances between pianos in the perceptual space constructed from the triadic comparisons. The listening experiments discussed in this chapter are an example of the "listening" approach shown in the schema of Figure 1.2.

In Chapter 4 the perceptual similarity among sounds is simulated using a computational model of the effective processing of the auditory system. The sounds are "presented" to the model in exactly the same way as in the instrument-in-noise test validated in the previous chapter. The simulated thresholds are significantly correlated with the experimental thresholds, when only a portion (onset) of the sounds is used as input to the model. These results suggest that the auditory cues available in the starting part of the sounds are sufficient to reach human performance with the model. The content of this chapter is an example of the "perceptual modelling" approach shown in the schema of Figure 1.2.

With the aim of broadening the use of the computational model of Chapter 4 to a different acoustic environment, in Chapter 5 the computational model is used to simulate the similarity of piano sounds in a reverberant condition. The reverberation is applied to the same piano sounds used in Chapters 3 and 4 by means of digital convolution. The effect of reverberation on the piano sounds introduces a moderate change in their relative position in the perceptual similarity space. The experimental results of the instrument-in-noise test as well as the simulated results from the computational model also account for this change.

In Chapter 6 a computational model (Processor E in Table 1.1) similar to that of the previous chapters is used to simulate the perceived reverberation of different orchestra instrument sounds in 8 different acoustic environments. The model is set-up in a binaural configuration and a different central processor is used to generate reverberance estimates. Experimental results for the same instrument sounds are provided. The reverberance estimates of the model for within-instrument conditions are correlated with the experimental results. This study case corresponds to an example of the "computational listening" approach shown in the schema of Figure 1.2.

In Chapter 7 the results and conclusions drawn from each chapter are briefly summarised. We discuss the context in which the auditory modelling approach was used, including perspectives for further research. This discussion is centred on further improvements that could be introduced to the auditory model and their possible implications in the unified computational framework.

2 Perceptual evaluation of instrument sounds using classic psychoacoustic descriptors[1]

2.1 Introduction

One way to better understand the properties of a musical instrument is to compare sound recordings of that instrument in controlled situations with synthesised sounds generated with physical models that recreate such situations. These sounds can be compared adopting a "computational listening" approach (see Figure 1.2 of the previous chapter). Since musical sounds are received and processed by the human hearing system, the comparison between sounds should be ideally based on perceptual criteria.

Studies in the field of psychoacoustics have addressed the problem of sound perception by developing (psychoacoustic) audio descriptors. As pointed out in the previous chapter (see Section 1.3.2), this development has been done by fitting algorithms of sound processing to experimental data obtained primarily with artificial test stimuli using the method of magnitude estimation (Stevens, 1955; Fastl, 1977; Zwicker, 1977; Kemp, 1982; Fastl, 1982, 1983; Daniel & Weber, 1997). These metrics have also been used to analyse other types of sounds such as speech, music, soundscapes, and sounds for product design (see, e.g., Terhardt, 1978; Genuit, 1997; Widmann, 1997; Yang & Kang, 2013).

In this chapter we compare recorded and synthesised sounds of an instrument called hummer, also known as the "voice of the dragon".

[1] This chapter is largely based on:
A. Osses, R. Kim, and A. Kohlrausch (2015). "Perceptual evaluation of differences between original and synthesised musical instrument sounds: the role of room acoustics". Proceedings of EuroNoise. C. Glorieux (Ed.), pp. 2561–2566. Maastricht, the Netherlands.
A. Osses, and A. Kohlrausch (2014). Perceptual evaluation of differences between original and synthesised musical instrument sounds. Actas 9th Iberoamerican Congress on Acoustics FIA. J. Arenas (Ed.) pp. 987–997. Valdivia, Chile.

The comparison is done using classic psychoacoustic metrics –loudness (loudness fluctuations), roughness, fluctuation strength– applied to hummer sounds available from a previous research project (Nakiboğlu et al., 2012; Hirschberg et al., 2013), where no quantitative evaluation of the agreement between their synthesised and recorded sounds was reported. Another motivation to evaluate hummer sounds is their simple nature: the sounds contain mainly one tonal component that oscillates periodically in frequency and amplitude (see panel B of Figure 1.1, page 5). Additionally, the envelope of the sounds is not perfectly regular, having a slowly-varying pattern in time. The aim of this chapter is, therefore, to compare available sounds of this simple musical instrument (recorded and synthesised) using quantitative evaluation criteria based on existing psychoacoustic metrics.

Since the evaluation criteria are based on applying the concepts of loudness, loudness fluctuations, roughness and fluctuation strength, we start the chapter by describing relevant aspects of these descriptors. In addition to these descriptors, F0 estimates are used to evaluate pitch variations in the test sounds. During the analysis, particular emphasis is given to the sensations of fluctuation strength and roughness. These descriptors characterise temporal fluctuations in amplitude and in frequency and are found naturally in everyday sounds.

2.2 Description of the method

The evaluation between sounds is done by comparing a number of features extracted from each of the sounds. To add a perceptual component, a set of psychoacoustic descriptors is used to extract those sound features. A summary of the descriptors used in this chapter is presented in Table 2.1. Further details are described in the subsequent sections.

Descriptors 1-2: Loudness and loudness fluctuations
Loudness corresponds to the perceptual correlate of the sound pressure level and is expressed in sone. The reference sound producing 1 sone is a 1-kHz sine tone with an SPL of 40 dB. A level increase of 10 dB leads roughly to a doubling of the loudness of a sound. In this chapter the loudness is obtained from the dynamic loudness model (DLM) (Chalupper & Fastl, 2002). This model provides loudness estimates as a function of time and frequency.

In order to appropriately describe the concept of loudness fluctuations, we need to introduce a more detailed description of the DLM model. The

Table 2.1: Summary of the psychoacoustic descriptors used in this chapter. Further details are given in the text. The range of values were taken from the literature. The JND values are related to the noticeable differences of the attributes in the range of the reference value. The JND for loudness was estimated considering an intensity-JND (ΔI) of 1 dB for a 1-kHz pure tone of 36 dB, as reported by Rabinowitz (1970). The JNDs for Roughness and Fluctuation strength were taken from Fastl and Zwicker (2007, their Chapters 10 and 11). The maximum values for fluctuation strength and loudness were taken from Fastl and Zwicker (2007, their Figures 10.2a and 16.1), and for roughness from Daniel and Weber (1997, their Figure 9).

Descriptor	unit	range	reference	JND
Loudness (N)	sone	$0-120$	1 sone	0.07 sone (ΔN= 7%)
Loudness fluctuation (L_G)	dB			$\Delta L_G \approx 1$ dB*
Roughness (R)	asper	$0-3.2$	1 asper	0.17 asper (ΔR = 17%)
Fluctuation strength (FS)	vacil	$0-3$	1 vacil	0.10 vacil (ΔFS = 10%)
Fundamental frequency (F0)	Hz		f_n Hz	ΔF0 $\approx$ 0.4%

(*)In this chapter we assumed that a difference of 1 dB at each critical-band level L_G as a function of frequency can be used as an estimate of the JND for loudness fluctuations.

block diagram of the model is shown in Figure 2.1. First, the incoming input signal is high-pass filtered ($f_{\text{cut-off}}$= 50 Hz). Then, an auditory filter bank consisting of 24 equidistant frequency bands with 1 Bark[2] distance is applied. The auditory bands have centre frequencies that range from 50 Hz (0.5 Bark) to 13500 Hz (23.5 Bark). In the "Envelope extraction" stage, the envelope of each auditory band is extracted by computing a short-term root-mean-square value. Main excitation patterns are obtained after accounting for the transmission from free-field through the outer and middle ears. This is obtained by applying an amplitude weighting a_0 as a function of frequency (see Fastl & Zwicker, 2007, their Figure 8.18). In the stage of "Loudness transformation" the excitation patterns are converted into main loudness by applying a compressive relation. This is followed by the (temporal) post-masking stage, where the effects of forward masking are accounted for. This is done by appending temporal tails onto the loudness patterns. Subsequently, an upward spread of masking is applied to the loudness patterns as a function of frequency at each time stamp. The resulting patterns are called specific loudness patterns. Finally, the patterns are integrated across frequency to obtain an instantaneous loudness estimate as a function of time. This temporal pattern is then smoothed in the "Temporal integration" stage by applying a low-pass filter (LPF) ($f_{\text{cut-off}}$= 8 Hz) to obtain the final "perceived" time-varying loudness.

[2]The critical-band rate z expressed in Barks corresponds to one of the frequency scales that is inspired by the frequency representation in the auditory system. A brief overview of this scale is given in Appendix A.

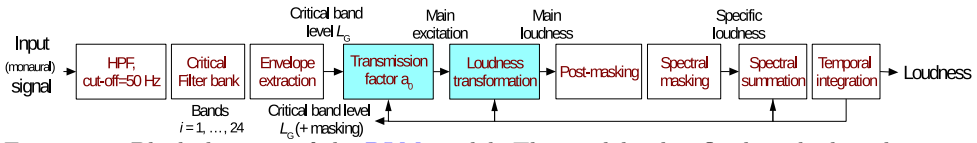


Figure 2.1: Block diagram of the DLM model. The model is briefly described in the text.

As an estimate of the loudness fluctuation of a sound, the critical-band levels L_G are used. They correspond to a representation of the envelope of the sound in dB as a function of frequency. In order to obtain critical band levels L_G that account for the temporal and spectral masking, the stages of "Loudness transformation" and "Transmission factor a_0" are reversed using the low-pass filtered specific loudness patterns. This is indicated in Figure 2.1 by the arrows in the lower part of the diagram. The reversed stages are highlighted in the diagram. The resulting L_G levels are labelled as "Critical band level L_G (+masking)" in the diagram. The minimum and maximum level patterns are estimated from the percentiles 5 and 95, respectively. Since the analysis presented in this chapter considers only "short signals" of 1.2 s (hummer sound, acoustic mode 2) or less, these percentiles are assessed over the entire duration of the sounds.

Descriptor 3: Roughness

Roughness (R) is a metric that describes how "rough" a sound is and is caused by the presence of rapid amplitude and/or frequency modulations with modulation rates between 15 and 300 Hz. The sensation of "roughness" has a bandpass characteristic with a maximum near the frequency of 70 Hz. Roughness is expressed in asper, where a sound producing 1 asper corresponds to a 1-kHz sine tone, 100% sinusoidally amplitude-modulated, with a modulation frequency of 70 Hz and an SPL of 60 dB (Kemp, 1982; Daniel & Weber, 1997). The lower limit of roughness perception is 0.07 asper and several authors agree that a relative variation of about 17% elicits a just-noticeable change in roughness (Vogel, 1975; Daniel & Weber, 1997; Fastl & Zwicker, 2007, Chapter 11). The model described by Daniel and Weber (1997) is used in this chapter. Particularly, we used the model outputs of main roughness and specific roughness.

Descriptor 4: Fluctuation strength

The metric of fluctuation strength (FS) is used to describe slow amplitude and/or frequency modulations with modulation rates below 20 Hz.

The sensation of fluctuation strength has a bandpass characteristic with a maximum around the frequency of 4 Hz. The range of modulations below 20 Hz has been shown to be of special interest for speech intelligibility (Drullman et al., 1994; Shannon et al., 1995) as well as for the perception of rhythm, which is related to the average syllable rate at amplitude modulations (AMs) of around 4 Hz (see, e.g., Leong et al., 2014). Fluctuation strength is expressed in vacil, where a sound producing 1 vacil corresponds to a 1-kHz sine tone, 100% sinusoidally amplitude-modulated, modulation frequency of 4 Hz and an SPL of 60 dB (Fastl, 1982, 1983). A relative variation of about 10% is believed to elicit a just-noticeable change in FS (Fastl & Zwicker, 2007, their Chapter 10). The model described by García (2015) and Osses et al. (2016) is used in this chapter. This model has been adapted from an algorithm used to assess roughness. The FS model is described in detail in Appendix B.

Descriptor 5: Fundamental frequency

The periodicity of a sound can be estimated by calculating the fundamental frequency (F0), which is expressed in Hz. F0 estimates are used to investigate the frequency variations of a given sound. For hummer sounds, these variations are related to Doppler shifts. In this context, the difference between the minimum and maximum F0 estimates ($\mathrm{F0_{range}} = \mathrm{F0_{max}} - \mathrm{F0_{min}}$) is used to evaluate the F0 range. For comparing F0 patterns as a function of time, the absolute difference between the F0 estimates of the test sounds (recorded and simulated sounds) normalised to the acoustic mode frequency f_n is used ($\Delta\mathrm{F0}[\%] = 100 \cdot \|\mathrm{F0_{rec}} - \mathrm{F0_{sim}}\|/f_n$). For sinusoidally frequency-modulated sounds ($f_{\mathrm{mod}} = 4$ Hz) varying by $\pm\Delta f$ around a carrier frequency f_c, just-noticeable changes in carrier frequency of 0.42% and 0.35% can be estimated for the frequencies of $f_2 = 424.4$ Hz and $f_4 = 851.8$ Hz (Fastl & Zwicker, 2007). These frequencies are of interest to evaluate hummer sounds because they correspond to its measured resonance frequencies in acoustic modes 2 and 4. F0 estimates are obtained using the Praat software (Boersma, 1993; Boersma & Weenink, 2001).

2.2.1 Comparing two sounds

The comparisons are based on the use of psychoacoustic descriptors. For each descriptor, test sounds differing by more than a minimum detectable change (one JND), are labelled as different enough to be distinguished from each other. The JNDs for each psychoacoustic descriptor are summarised in Table 2.1.

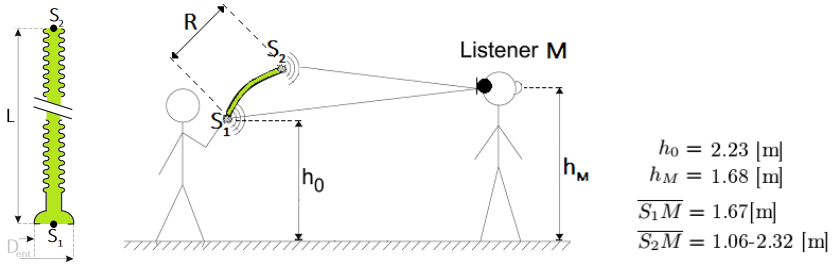


Figure 2.2: Schematic drawing of a hummer. The hummer has a length L of 70 cm, the inlet (S_1) has an entrance diameter D_{ent} of 3.3 cm. The opposite end of the hummer is identified as the outlet (S_2). Note that the distances in this drawing are not to scale. Some pictures of the hummer can be found in the study by Hirschberg et al. (2013). This figure was adapted from Nakiboğlu et al. (2012).

2.3 Study case: Comparison between recorded and synthesised hummer sounds

2.3.1 Principle of sound generation

The hummer is a flexible plastic corrugated pipe with both ends open. A schematic geometry of the hummer and typical dimensions are shown in Figure 2.2. To generate sound, the hummer has to be rotated at a certain speed in order to excite the natural frequencies of the pipe. The resonance frequencies f_n of the system as a function of the acoustic mode n are given by:

$$f_n \approx n \cdot \frac{c_{\text{eff}}}{2L} \quad \text{with } n = 2, 3, \ldots \tag{2.1}$$

where c_{eff} corresponds to the effective speed of sound in the tube and L corresponds to the length of the pipe. The effective speed of sound is approximately 310 m/s (Nakiboğlu et al., 2012). The resonance frequencies f_n are shown in Table 2.2. The theoretical frequencies f_n can be obtained using Equation 2.1. The "measured" frequencies were derived from the sound recordings.

The rotational movement of the hummer produces a periodic variation in distance between sound source and listener, which leads to positive and negative frequency shifts due to the Doppler effect. This variation is related to the rotation period of the hummer.

2.3.2 Stimuli

In this section a brief description of the existing recordings and the synthesised hummer sounds is presented. More detailed information about

Table 2.2: Resonance frequency f_n and rotation period Ω_n for the hummer at different rotation speeds (modes 2 and 4) derived from both, theory (Equation 2.1) and the recordings.

Acoustic mode n	Frequency f_n [Hz] Theory	Measured	ΔF0 [%]	Period Ω_n [s]
2	442.9	424.4	4.2	0.602
4	885.7	851.8	3.8	0.296

the mechanical measurement set-up used for the sound recordings is given by Hirschberg et al. (2013). The physical model used for synthesising the hummer sounds is described by Nakiboğlu et al. (2012).

Recorded sounds

The recordings were made using a mechanical set-up, where the hummer was attached to a bicycle wheel with an adjustable rotation speed. The set-up was installed in a semi-anechoic room (volume of 100 m^3) that had a non-reflecting floor. The resulting environment was nearly anechoic. This means that the microphone M captured only contributions from the sources S_1 and S_2. Figure 2.2 gives a schematic view of the position of the hummer with respect to the microphone M. The mechanical system on which the hummer was mounted is not shown in the figure.

The hummer was attached to the spikes of a 26" bicycle wheel. The inlet S_1 was placed close to the axis of rotation (wheel axis). The outlet S_2 was at a distance of 0.70 m from the wheel axis, approximately 0.30 m outside the radius of the wheel. The wheel was mounted on a structure (oriented horizontally), at a height of 2.23 m above the floor. The wheel axis was defined to be at coordinates (0,0,2.23) m.

A microphone B&K type 4190, located at (1.58, 0, 1.68) m, was used to record the hummer. The microphone was located, thus, at a distance of 1.67 m from the centre of rotation. Each recording had a duration of 20 s and was sampled at 10 kHz, with an amplitude resolution of 16 bits. The measured resonance frequencies differed by about 4% from the approximation given by Equation 2.1, as shown in Table 2.2.

The recorded signals were re-sampled at 44.1 kHz, with an amplitude resolution of 16 bits. The average level was adjusted according to the reference levels of 54 and 72 dB SPL at 1.67 m from the origin of the system for the acoustic modes 2 and 4, respectively.

The waveforms of the recorded hummer signals as used in this chapter are shown in panel A of Figure 2.3. As a consequence of the movement

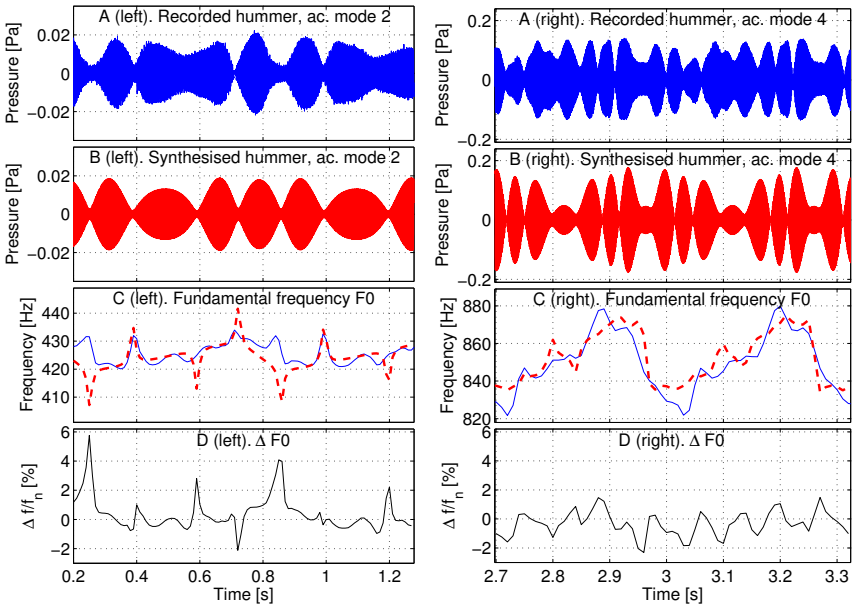


Figure 2.3: Hummer sounds in the acoustic mode 2 (left panels) and 4 (right panels). In panels A and B, the recorded and synthesised waveforms are shown, respectively. In panel C, F0 estimates obtained using the autocorrelation-based F0 extractor available in the software Praat are shown. In panel D, the differences [%] between F0 estimates are shown relative to $f_2 = 424.4$ Hz (in mode 2) and $f_4 = 851.8$ Hz (in mode 4).

of S_2, the hummer sounds present a Doppler shift around their natural frequency f_n, as shown in panel C (solid blue line) of the figure.

The mechanical system produced an audible noise in the recordings which is not present in the synthesised sounds. For this reason, in the comparison between recorded and synthesised sounds, only those frequency components that are around f_n are considered. In acoustic mode 2 (f_2 of 424.4 Hz), the analysis considered all frequency components between 300 Hz (2.9 Bark) and 1000 Hz (8.5 Bark). In acoustic mode 4 (f_4 of 851.8 Hz), the analysis considered all frequency components between 650 Hz (6 Bark) and 1400 Hz (10.7 Bark).

Synthesised sounds

Considering a hummer of length $L = 0.7$ m, as represented in Figure 2.2, the instrument can be modelled as two monopole sound sources. The inlet, located near the axis of the wheel, with an entrance diameter of $D_{\text{ent}} = 3.3$ cm, was modelled as a fixed source S_1, while the outlet was modelled as a rotating source S_2 with a rotation period of Ω_n. Because of the flexible nature of the hummer, an effective rotation radius R of 0.67 m was used.

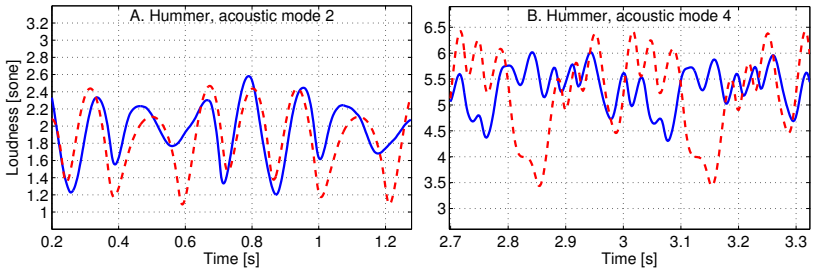


Figure 2.4: Loudness of recorded (solid) and synthesised (dashed) hummer signals in the anechoic condition for the acoustic modes 2 (panel A) and 4 (panel B). Only the loudness contribution of frequency components between $z_{\min}$ and $z_{\max}$ were taken into account.

The synthesised waveforms were obtained using the physical model described by Nakiboğlu et al. (2012). The model accepts L, D_{ent}, R, Ω_n, f_n, the parametrised positions of the sound sources $S_{1,2}(t)$, and the listener (microphone) location as input parameters. The measured resonance frequencies f_n and rotation periods Ω_n presented in Table 2.2 were used instead of their theoretical values.

The synthesised sounds were sampled at 44.1 kHz with an amplitude resolution of 16 bits. The average level was adjusted according to the reference levels of 54 and 72 dB SPL at 1.67 m from the origin of the system for the acoustic modes 2 and 4, respectively. The waveforms of the synthesised hummer signals are shown in panel B of Figure 2.3. The shift in F0 caused by the movement of S_2 is indicated by the red dashed lines in panel C of Figure 2.3.

2.4 Results

The following results were obtained using two rotation periods of the hummer signals. For recorded sounds, the most stable periods were chosen.

2.4.1 Loudness

The results for the loudness estimates as a function of time (output of the DLM model) are shown in Figure 2.4. The minimum, median, and maximum loudness values were assessed as the percentiles L_5, L_{50} and L_{95}, respectively, and they were obtained by performing the spectral summation and temporal integration of the specific loudness patterns within a frequency range around the F0 of each mode. Those loudness values are shown in Table 2.3. The loudness difference ΔL_{50} in acoustic mode 2 was $\|2.0 - 1.9\| = 0.1$ sone, while the same loudness value was obtained in mode 4: $\|5.4 - 5.4\| = 0$. These values differ by approximately

Table 2.3: Summary of the specific loudness patterns in percentiles for 2 periods of rotation of the hummer signals. Percentile 5 and 95 represent minimum and maximum values, respectively. Percentile 50 is an estimate of the mean loudness value. To assess these values, only the frequency components in the range $(z_{\min}, z_{\max})$ were taken into account.

Acoustic Mode n / Type	Frequency limit [Bark] $z_{\min}$-$z_{\max}$	Loudness [sones] L_5	L_{50}	L_{95}	$L_{95} - L_5$
2 / recorded	2.9 - 8.5	1.3	2.0	2.5	1.1
2 / synthesised	2.9 - 8.5	1.2	1.9	2.4	1.2
4 / recorded	6.0 - 10.7	4.5	5.4	5.9	1.4
4 / synthesised	6.0 - 10.7	3.7	5.4	6.3	2.7

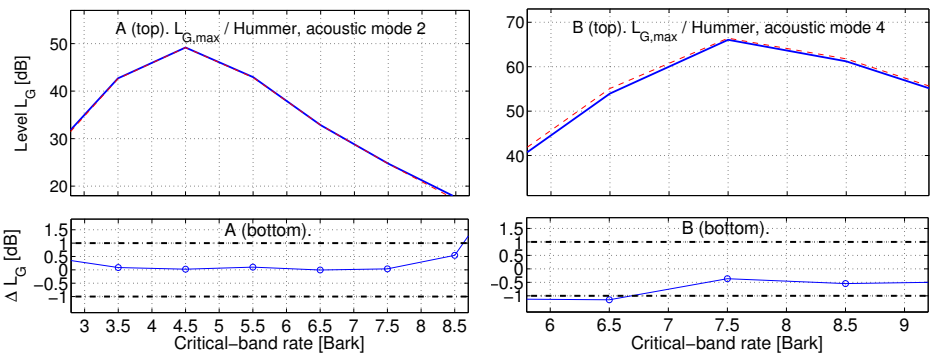


Figure 2.5: Maximum critical-band levels $L_{G,\max}$ for recorded (blue solid line) and synthesised (red dashed line) hummer signals in the acoustic modes 2 (panel A) and 4 (panel B). In the bottom panels, the differences between the recorded and synthesised signals are shown. The black dashed-dotted lines indicate the assumed JND of 1 dB.

one JND or less. Although the reported JND for a 40-dB tone presented in Table 2.1 is 0.07 sone, the JND for higher levels increases to 0.12 sone at 54 dB SPL (4.6% of relative change) and to 0.30 sone at 72 dB SPL (3.3% of relative change). If we consider a positive difference to be attributed to higher values in the recorded signals, then in mode 2, the minimum L_5 and maximum L_{95} estimates have a good agreement with a deviation $\Delta L_5 = 1.3 - 1.2 = 0.1$ sone and $\Delta L_{95} = 2.5 - 2.4 = 0.1$ sone, which is still within the range of one JND. Although in acoustic mode 4, the synthesised signal is as loud as the recorded signal ($\Delta L_{50} = 0$), its maximum value is higher ($\Delta L_{95} = 5.9 - 6.3 = -0.4$ sone) and its minimum is lower ($\Delta L_5 = 4.5 - 3.7 = 0.8$ sone). The underestimation of the minimum loudness values ($\Delta L_5 = 0.8$ sone > 1 JND), is particularly visible in panel B of Figure 2.4, where the loudness of the synthesised sound has minimum values of nearly 3.4 sone at 2.86 and 3.16 s, while the recorded signal has a minimum value of about 4.3 sone.

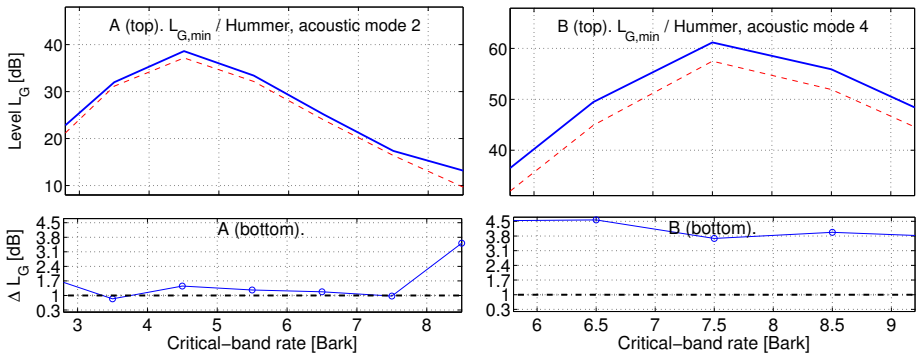


Figure 2.6: Minimum critical-band levels $L_{G,\min}$ for recorded (solid) and synthesised (dashed) hummer signals in the acoustic modes 2 (panel A) and 4 (panel B). In the bottom panels, the differences between the recorded and synthesised signals are shown. In panel A, the differences are slightly larger than one JND, with a more pronounced difference above 7.5 Bark (853 Hz). In panel B, the $L_{G,\min}$ levels of the synthesised signals are always below the levels of the recorded signals, with an underestimation that reaches 4.6 dB at 6.5 Bark (720 Hz). The assumed JND of 1 dB is indicated by the black dashed-dotted line.

2.4.2 Loudness fluctuations

The results for the critical-band levels L_G are shown in Figures 2.5 and 2.6. The maximum critical-band levels $L_{G,\max}$ as a function of frequency can be used as an estimate of the maximum masking pattern produced by a signal. Likewise, the minimum critical-band level $L_{G,\min}$ can be used to estimate minimum masking patterns. The $L_{G,\max}$ levels of recorded and synthesised hummer signals are shown in Figure 2.5. The levels differ by less than 1 dB for signals in acoustic mode 2 (panel A of the figure). For signals in mode 4 (panel B of the figure), the synthesised sound has slightly overestimated loudness fluctuation values for frequencies below 6.7 Bark (740 Hz), producing a $\Delta L_{G,\max}$ of -1.1 dB at 6.5 Bark (720 Hz). This means that these level differences are likely to be perceived for frequencies below 6.7 Bark (740 Hz), where the JND is just exceeded.

The differences were larger in the minimum masking patterns, shown in Figure 2.6. For both acoustic modes the synthesised signals had a $L_{G,\min}$ pattern below those of the recorded signals. In mode 2 (panel A of the figure) the differences were equal to or lower than 1.5 dB for frequency components between 3.0 Bark (313 Hz) and 7.7 Bark (880 Hz). The differences were larger for the masking patterns in mode 4 (panel B of the figure) where synthesised signals produced $L_{G,\min}$ levels that are lower by at least 3.7 dB. This means that for both modes, the differences between synthesised and recorded hummer signals are likely to be perceived. The differences are however more prominent in acoustic mode 4

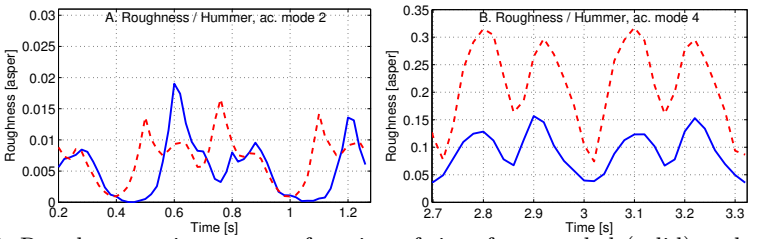


Figure 2.7: Roughness estimates as a function of time for recorded (solid) and synthesised (dashed) hummer signals. The hummer signals in the acoustic mode 2 (panel A) do not produce any sensation of roughness (R< 0.07 asper). In the acoustic mode 4 (panel B), the recorded signal has an overall R value which is just above threshold of 0.08 asper, while the synthesised sound has a higher sensation, with an overall R value of 0.22 asper.

($\Delta L_{G,\min} \geq 3.7$ dB) than in mode 2 ($\Delta L_{G,\min} \leq 1.5$ dB for frequencies below 7.7 Bark).

2.4.3 Roughness

The results for the R estimates as a function of time are shown in Figure 2.7. The results for the (overall) specific roughness R_{spec} patterns as a function of frequency are shown in Figure 2.8. The results for the hummer signals in acoustic mode 2 (panel A in Figures 2.7 and 2.8) have an R value below the minimum audible threshold of 0.07 asper, meaning that the signals do not elicit any roughness sensation. In mode 4 (panel B of the figures), the recorded signal (blue solid line) has an overall R value which is just above threshold of 0.08 asper with minimum and maximum values of $R_5 = 0.04$ asper (below threshold) and $R_{95} = 0.15$ asper, while the synthesised sound (red dashed line) has a higher sensation, with an overall R value of 0.22 asper and minimum and maximum values of $R_5 = 0.08$ asper and $R_{95} = 0.31$ asper. The JND value for a roughness of 0.22 asper is 0.04 asper (17% of 0.22 asper). Hence, the synthesised signal produces a roughness sensation that is markedly higher to that produced by the recorded signal ($R_{sim} - R_{rec} = 0.22 - 0.08$ asper $= 0.14$ asper > 0.04 asper). Although the signals in mode 2 do not produce any sensation of roughness and the recorded hummer sound in mode 4 is just above the roughness threshold, all four R_{spec} patterns in Figure 2.8 have a maximum value at the critical bands with centre frequencies closer to the F0s of the respective modes. In mode 2, the maximum occurs in the band centred at 4.0-4.5 Bark (close to $f_2 = 4.1$ Bark $= 424.4$ Hz). In mode 4, the maximum occurs in the band centred at 7.5 Bark (852.7 Hz, close to $f_4 = 851.8$ Hz).

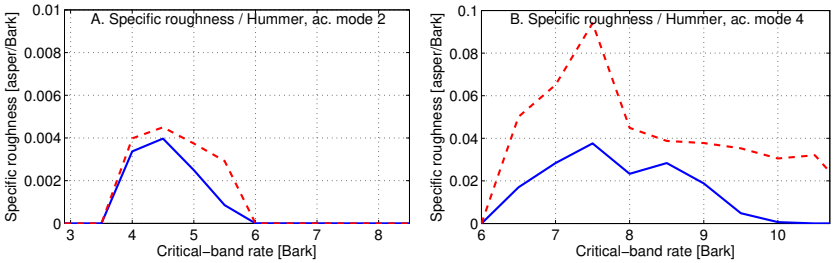


Figure 2.8: Average specific roughness patterns R_{spec} for recorded (solid) and synthesised (dashed) hummer signals. All four R_{spec} patterns have a maximum value at the critical bands with centre frequencies closer to the F0s of the respective modes. In acoustic mode 2 (panel A), the maximum occurs in the band centred at 4.0-4.5 Bark (417.3-473.4 Hz, close to $f_2 = 424.4$ Hz). In acoustic mode 4 (panel B), the maximum occurs in the band centred at 7.5 Bark (852.7 Hz, close to $f_4 = 851.8$ Hz).

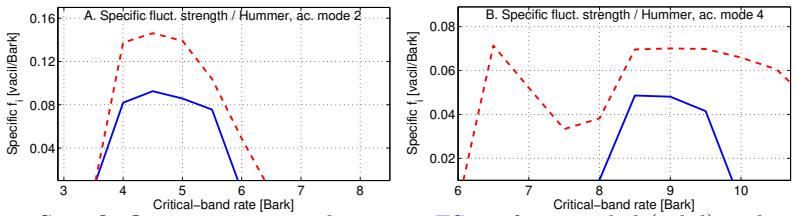


Figure 2.9: Specific fluctuation strength pattern FS_{spec} for recorded (solid) and synthesised (dashed) hummer signals. The overall FS values that can be obtained by integrating the area under the FS_{spec} patterns are 0.18 and 0.29 vacil for the recorded and synthesised signals in acoustic mode 2, and 0.07 and 0.30 vacil in acoustic mode 4.

2.4.4 Fluctuation strength

The results for the patterns of specific fluctuation strength (FS_{spec}) are shown in Figure 2.9. For this analysis, 2-s section of recorded and synthesised hummer sounds were used as input to the FS model. The analysis window of the model was set to 2 s, meaning that the algorithm only returned one overall FS value and one pattern of specific fluctuation strength FS_{spec}. The overall FS values for recorded and synthesised signals in acoustic mode 2 were 0.18 vacil and 0.29 vacil, respectively. The FS values for the signals in acoustic mode 4 were 0.07 vacil and 0.30 vacil. In both modes the synthesised hummer signals elicit a higher sensation of fluctuation than those of the recorded signals and they differ by more than one JND. The JNDs for the FS values of 0.29 and 0.30 vacil are about 0.03 vacil. Therefore, the differences are $FS_{sim} - FS_{rec} = 0.29 - 0.18 = 0.11$ vacil > 0.03 vacil in mode 2, and $FS_{sim} - FS_{rec} = 0.30 - 0.07 = 0.23$ vacil > 0.03 vacil in mode 4, i.e., in both modes the differences in FS are larger than one JND. The FS value of the recorded hummer signal in the acoustic mode 4 is very low (0.07 vacil) and, therefore, it can be labelled as a non-fluctuating sound.

2.4.5 Fundamental frequency

The results for the F0 estimation of recorded (blue line) and synthesised sounds (red dashed line) are shown in panel C of Figure 2.3, where a pitch estimate was found for every audio segment[3]). In acoustic mode 2, the F0 estimates for the recorded signals vary between 420 and 434 Hz ($\mathrm{F0_{range}} = \mathrm{F0_{max}} - \mathrm{F0_{min}} = 14$ Hz), while the estimates for the synthesised signals vary between 407 and 442 Hz ($\mathrm{F0_{range}} = 35$ Hz). The F0 patterns are periodic, following the rotation period of the hummer of about 0.6 s ($f_{\mathrm{rot}} = 1.7$ Hz). In acoustic mode 4, the F0 estimates for the recorded signals vary between 822 and 878 Hz ($\mathrm{F0_{range}} = 56$ Hz), while for the synthesised signals they vary between 835 and 874 Hz ($\mathrm{F0_{range}} = 39$ Hz). The F0 patterns in this mode have a rotation period of about 0.3 s ($f_{\mathrm{rot}} = 3.3$ Hz). The differences between F0 estimates (normalised to f_n) are shown in panel D of Figure 2.3. In mode 2 (panel D, left), the ΔF0 ranges from -2.1% to 5.8%, with an unsigned average of 0.7%. In mode 4 (panel D, right), the ΔF0 ranges from -2.3% to 1.5%, with an average of 0.7%. The average differences in both modes exceed the reported JNDs for variations in frequency of stationary FM tones (0.42% and 0.35%, respectively).

2.5 Discussion

The results of the comparison between recorded and synthesised hummer signals are summarised in Table 2.4. The synthesised hummer sounds showed a higher similarity[4] with the recorded signals in mode 2 than in mode 4. In mode 2, differences that are unlikely to be perceived were found for the descriptors of loudness, loudness fluctuation ($L_{G,\mathrm{max}}$), and roughness. The descriptors of loudness fluctuation ($L_{G,\mathrm{min}}$), fluctuation strength, and F0 indicated that perceptual differences between the synthesised and recorded sounds exist[5]. In mode 4, differences between the recorded and synthesised signals that are likely to be perceived were found for the descriptors of loudness ($L_{95} - L_5$), loudness fluctuation ($L_{G,\mathrm{min}}$), roughness and fluctuation strength. The discussion presented

[3] Pitch estimates were obtained for 40-ms segments with a hop-size of 10 ms and F0 candidates between 75 and 1400 Hz. The frequency contours were obtained in the Praat software using the following command: To pitch (ac)... 0.01 75 15 no 0.01 0.45 0.01 0.35 0.14 1400.

[4] The term similarity is used here to refer to sounds that are not distinct enough according the selected psychoacoustic descriptors.

[5] As pointed out in Table 2.4, the differences in minimum loudness fluctuation and F0 are not much larger than the assumed JNDs. It is therefore unclear whether the use of more accurate JNDs (assessed for hummer signals) may still have lead to perceptible differences. For instance, for F0 differences the actual JND should be larger than the assumed JND, because the hummer has a dynamic variation (Doppler shift) while the assumed JND is valid for stationary FM tones.

Table 2.4: Summary of the comparison between synthesised and recorded hummer signals.

Descriptor		Are the hummer signals "different"? Mode 2	Mode 4	Figure Nr.
Loudness	ΔL_{50}	No	No	2.4
	$L_{95} - L_5$	No	Yes	2.4
Loudness fluctuation	$\Delta L_{G,\max}$	No	Yes	2.5
	$\Delta L_{G,\min}$	Yes*	Yes	2.6
Roughness	ΔR	No**	Yes	2.7
Fluctuation strength	ΔFS	Yes	Yes	2.9
Fundamental frequency	ΔF0	Yes*	Yes*	2.3

(*)The differences found for $\Delta L_{G,\min}$ patterns (in mode 2) and ΔF0 were not much larger than the assumed JNDs. The assessment of experimental JNDs may reveal whether these differences are actually perceptible. (**)The hummer signals in mode 2 did not elicit roughness.

next is focused on an analysis of the descriptors of roughness and FS. An analysis based on these descriptors allows the description of sounds in terms of their amplitude and frequency variations, which are prominent characteristics of the hummer signals.

2.5.1 Roughness

The hummer signals in acoustic mode 2 had R estimates below its minimum audible threshold of 0.07 asper. This means that the amplitude modulations (amplitude envelope) of the hummer signals have a periodicity that is not fast enough to enter the frequency range that elicits a sensation of roughness. This is also the case for their frequency modulations. The repetition rates of the frequency modulations follow the frequency of rotation of the hummer, which are 1.7 Hz (for $\Omega_n = 0.602$ s) and 3.3 Hz (for $\Omega_n = 0.296$ s) for the signals in modes 2 and 4, respectively. Both rates are below 20 Hz. Hence, the audible R values found for the signals in mode 4 should only be caused by their amplitude variations. Let us focus on the synthesised hummer sound in mode 4, which presents the highest R estimates. Its waveform, which is replotted in panel A of Figure 2.10 (taken from panel B of Figure 2.3), has pronounced amplitude modulations, with a Hilbert envelope that has 8 local maxima within a period (black circle markers). These maximum values range between 67.8 dB (0.049 Pa) at the points marked as 4 and 8, and 78.6 dB (0.170 Pa) at the points marked as 1 and 7. It can be noted that the amplitude modulations that lead to the lower amplitude maxima (points marked as 4 and 8 in the figure) are found when the F0 estimates cross the nominal mode frequency. This happens when the moving source S_2 is either facing (S_2 at $[0.67, 0, 2.23]$ m) or opposing (S_2 at $[-0.67, 0, 2.23]$ m) the recording microphone. Let us now consider two

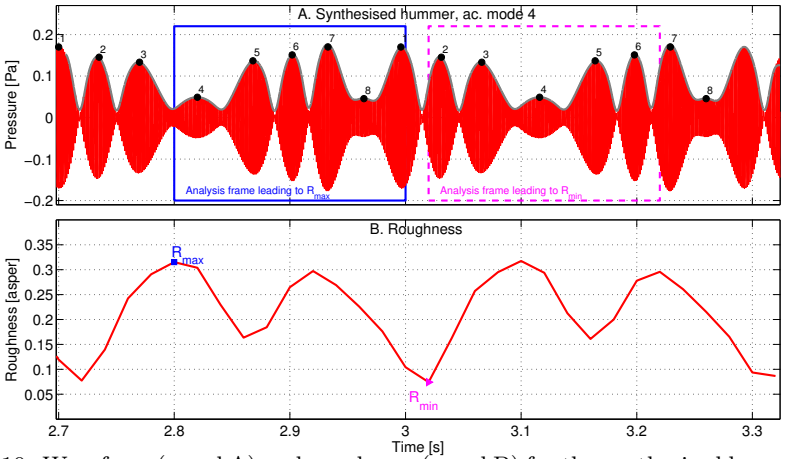


Figure 2.10: Waveform (panel A) and roughness (panel B) for the synthesised hummer sound in acoustic mode 4. Panels A and B are replotted from Figures 2.3 and 2.7, respectively. The waveform is shown together with its Hilbert envelope (grey thick line). Local maximum values of the envelope of the signal are indicated by dark circle markers and they are enumerated (1 to 8) in two periods of the hummer signal.

of the points at which R_{min} and R_{max} values occur, for instance, at 3.02 s (R= 0.07 asper) and 2.80 s (R= 0.32 asper), respectively. The R_{max} value is obtained considering the waveform samples between 2.80 and 3.00 s, as indicated by the blue rectangle in panel A of Figure 2.10. This analysis frame contains the two lower amplitude modulations (points 4 and 8) while the R_{min}-analysis frame (3.02-3.22 s, magenta dashed rectangle in the figure) contains only one (point 4). The presence of two lower amplitude modulations within one analysis frame seems to be enough to elicit a roughness sensation at their inherent modulation frequency around 25 Hz (duration between two consecutive minima of about 40 ms). It is important to emphasise that the elicited overall R of 0.32 asper (0.22 asper for the recorded hummer) is audible, but is still located in the lower end of the roughness scale. This means that the sensation of roughness is perceptible but not very prominent in the hummer sounds.

2.5.2 Fluctuation strength

Differences in acoustic mode 2

As just discussed, the differences between hummer sounds can be either attributed to amplitude or frequency modulations. For the signals in mode 2, estimates of loudness and maximum loudness fluctuation ($\Delta L_{G,\max}$) between recorded and synthesised sounds did not differ considerably in our analysis, while there was a slight underestimation of the

minimum loudness fluctuation observed in the synthesised signal, with an overall $\Delta L_{G,\min}$ of -1.5 dB (0.5 dB beyond the assumed JND). If we use loudness estimates as indicative of variations in the amplitude envelope, disregarding the 0.5-dB underestimation in $L_{G,\min}$, the difference in FS between hummer sounds should be caused by differences in their frequency modulations. The synthesised sound was found to have an F0 with a larger variation ($F0_{range}$) than that of the recorded sound (see Figure 2.3, panel C, left). The F0 estimates have a periodicity related to the rotation frequency of the hummer, in this mode of $f_{rot} = 1.7$ Hz ($\Omega_n \approx 0.6$ s). Since this frequency lies within the range of frequencies that are relevant for fluctuation strength, we may attribute the higher FS of the synthesised signal to its more prominent Doppler shift (higher $F0_{range}$ value) with respect to the recorded signal.

Differences in acoustic mode 4

For the signals in mode 4, the descriptors of loudness and loudness fluctuations already showed an underestimation of the minimum amplitude values. This means that at least part of the difference ($FS_{sim}-FS_{rec}=$ 0.23 vacil) between FS values can be attributed to amplitude modulations. The recorded and synthesised hummer sounds were found to have $F0$ ranges of 56 Hz ($\Delta f \approx \pm 28$ Hz) and 39 Hz ($\Delta f \approx \pm 20$ Hz), respectively. In an analysis presented in Appendix B, FM tones with a similar carrier frequency ($f_c = 851.8$ Hz), frequency deviation ($\Delta \pm 25$ Hz), modulation frequency ($f_{mod}= 4$ Hz), and no amplitude modulation (flat envelope) elicited FS model estimates of 0.11 vacil or less. Since the frequency modulations (FMs) follow a rotation frequency of $f_{rot} = 3.3$ Hz (close to $f_{mod} = 4$ Hz), the analysis shown in the appendix can be used to argue that the difference between FS estimates in mode 4 is unlikely to be produced by differences in the frequency modulation of the hummer sounds.

2.6 Conclusions

The methods presented in this chapter have been applied to recorded and synthesised sounds of an instrument called hummer. The analysis was based on five descriptors –loudness, loudness fluctuations, roughness, fluctuation strength, fundamental frequency–, that can be interpreted as an evaluation based on 5 dimensions. Within each of these dimensions, the psychoacoustic estimates obtained from the recorded and synthesised sounds were considered as similar if they differed by less than one JND

and as perceptually different otherwise. The results showed that the synthesised sounds are more similar to the recorded ones in acoustic mode 2, where two of the descriptors differed by less than one JND (loudness and roughness) and one descriptor was just above the JND (loudness fluctuation), than in mode 4, where only one of the descriptors met such a criterion (loudness, L_{50}).

The evaluated sounds are periodic and harmonic and they are characterised by the presence of both amplitude and frequency modulations. Based on these properties we assumed that the selected descriptors were appropriate to evaluate differences between recorded and synthesised hummer sounds. Other musical instruments may have properties that require another set of descriptors, which can increase the difficulty of the evaluation if more descriptors are needed, requiring more knowledge about the underlying JNDs. Some other instrument properties may be: (1) the presence of temporal transients, and; (2) the transition in pitch percepts from harmonic to non-harmonic segments within the sound.

In order to introduce the analysis of sounds that have temporal transients, recorded piano sounds are studied in the next chapters (Chapters 3, 4, 5). There, the perceptual similarity between sounds is approached as an experimental (performance) task and it does not require an a priori knowledge about the dimensions that are to be evaluated, as it was the case in this chapter.

3 Measuring the perceived similarity of instrument sounds using an instrument-in-noise test

In this chapter the comparison between sounds is approached as a discrimination task. This discrimination task has been adapted to assess the perceptual similarity of two test sounds. In the previous chapter, two sounds were "judged" as very similar if a given psychoacoustic metric provided values that differ by less than one JND. This situation would be comparable to a listening condition of the same two sounds with a level difference that is below the discriminability threshold.

In contrast to the use of a specific psychoacoustic metric, the proposed method is developed under the idea that, when comparing two sounds, a listener will use all available sound properties –or prominent features– rather than using a single property. The experiment is implemented as an "instrument"-in-noise task. The two sounds being evaluated are presented with an added specific noise. By adjusting the SNR in the course of the experiment the difficulty of the sound discrimination is manipulated. Two sounds that are similar will tolerate a low level of added noise (high SNR) to correctly discriminate one from the other in contrast to the case of two sounds that are more dissimilar, where a higher amount of noise (lower SNR) will be tolerated before the discriminability performance decreases. In other words, a strong correlation between SNR and similarity is expected. To produce this effect, however, the noises need to have similar spectro-temporal properties to those of the test stimuli. For that purpose the algorithm of the ICRA noises in speech has been adapted. A description to use this algorithm in the evaluation of a set of test stimuli is given. As study case, the instrument-in-noise test is used to evaluate recordings of one note played on seven Viennese pianos. The suggested method is compared to the method of triadic comparisons in a similarity assessment task.

3.1 Introduction

Perceptual similarity between elements is a problem approached in several disciplines and is normally assessed experimentally. Popular experimental tasks used to compare sounds are the method of triadic comparisons (Levelt et al., 1966; Fritz et al., 2010; Novello et al., 2011), pairwise comparisons (Grey, 1977; Grey & Gordon, 1978; Raake et al., 2014; Tahvanainen et al., 2015), free verbalisation rating, and categorisation (Dubois, 2000; Guastavino & Katz, 2004; Saitis et al., 2013). A review of these and other methods used in auditory research in the context of musical instruments is provided by Fritz and Dubois (2015) and also in the introduction of this thesis (Section 1.3). For the methods of triadic and pairwise comparisons, matrices indicating the preferences of the participants can be constructed. To further process the data, the preference matrices are normally converted into a mathematical space where the elements under test can be compared to each other. Techniques as MDS (Shepard, 1962; Kruskal, 1964b) and the use of the Bradley-Terry-Luce (BTL) scale (Bradley, 1953; Wickelmaier & Schmid, 2004) are examples of algorithms that allow such a comparison.

Despite all those experimental procedures to evaluate similarity, our interest is not only on knowing which sounds are more or less similar among each other but also on obtaining a quantifiable measure of those distances. In this chapter we show a way to reach that objective by conducting a listening test to discriminate two sounds using a 3-AFC experiment in noise, where the noise allows to change the similarity of the sounds being tested. In the next section the discrimination test or "instrument"-in-noise test is explained, providing a detailed explanation of the noise generation. As study case, a comparison of one note ($C\#_5$) of seven Viennese pianos from the 19^{th} century is given. A description of the method of triadic comparisons is also included. The triadic comparison test is used as reference method in the validation of the instrument-in-noise task.

3.2 Description of the method

A method to quantify the perceptual differences between sounds is presented in this section. The sounds are compared pairwise and they are embedded in a background noise at different SNRs. The method was developed under the rationale that two very different sounds must be easy to discriminate while two similar sounds must represent a more difficult

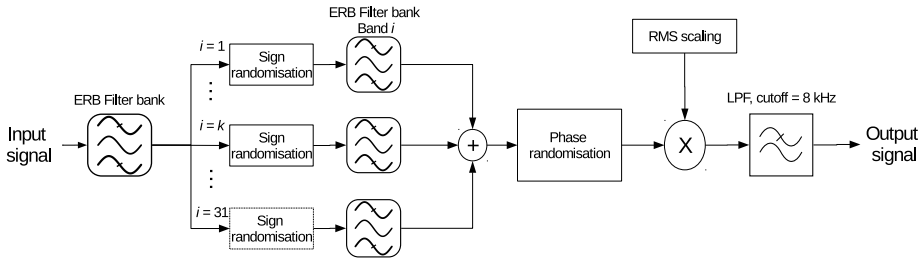


Figure 3.1: The principle of the ICRA noise generation, version A. For details in the procedure, refer to steps 1 to 6 in the text.

task. The similarity between two sounds within a trial is changed by presenting the sounds simultaneously with a specific noise. When the test sounds are more different, more noise (lower SNR) is tolerated until both sounds become undistinguishable. To deliver such results, however, the noise has to be carefully generated. The noise needs to have similar spectro-temporal properties to those of the test sounds. In the context of speech perception, the International Collegium of Rehabilitative Audiology (ICRA) developed an algorithm to generate random noises with such acoustic properties (Dreschler et al., 2001). We modified that algorithm to produce a suitable weighting of the properties of a musical instrument. The piano was chosen to exemplify the instrument-in-noise procedure. This choice was motivated by the strongly varying temporal properties and rich spectrum of the piano sounds.

3.2.1 Modified ICRA noise, version A

The procedure to generate the ICRA noises (version A[1]) introducing a "musical-instrument weighting" is shown in Figure 3.1 and can be summarised as follows:

1. **Band-split filter**: an input signal (musical instrument sound) is fed into a Gammatone filter bank. The Gammatone filter bank consists of 31 bands with centre frequencies between 87 Hz (3 ERB_N[2]) and 7820 Hz (33 ERB_N), spaced at 1 ERB. The all-pole Gammatone filter bank with complex outputs (only the real part is further processed) available in the Auditory Modelling Toolbox (AMT) for MATLAB was used for this purpose (Søndergaard & Majdak, 2013). The filter design and processing

[1] In a later stage of our research project, a second modification of the ICRA algorithm (version B) was developed. Version B of the ICRA algorithm is described and used in Chapter 5.

[2] The equivalent rectangular bandwidth (ERB) rate scale corresponds to one of the frequency scales that is inspired by the frequency representation in the auditory system. A brief overview of this scale is given in Appendix A.

introduced in this stage is equivalent to the "frequency analysis" stage described by Hohmann (2002).

2. **Sign randomisation**: the sign of each sample of the 31 filtered signals is either reversed or kept unaltered with a probability of 50% (multiplication by 1 or -1) (Schroeder, 1968). As a consequence of this process, the resulting waveforms have a flat spectrum while keeping the same temporal envelope characteristics and the same band level.

3. **Re-filtering per band-split filter**: the resulting signal from band i is fed into the ith band of the Gammatone filter bank. The index i represents each of the 31 bands.

4. **Add signals together**: the 31 filtered signals are added together.

5. **Phase randomisation**: the phase of the signal is randomised following a uniform distribution between 0 and 2π, this is done in the frequency domain by overlapping/adding the segments after an IFFT with a 87.5% overlap. The resulting signal is adjusted to have the same total RMS level as the input to the band-split filter stage.

6. **Low-pass filter at 8200 Hz**: an eight-order Butterworth filter with a cut-off frequency at the upper limit of the highest critical band ($f_{\text{cut-off}}$ at 8200 Hz$\approx$ 33.5 ERB_N) is applied. This filter is introduced to reduce undesired high frequencies as a consequence of the phase randomisation.

One fundamental change in the ICRA-noise algorithm compared to the original description by Dreschler et al. (2001) is the use of the 31-band Gammatone filter bank instead of the original band-split filter with cross-over frequencies at 800 and 2400 Hz, i.e., a LPF with cut-off frequency at 800 Hz, a band-pass filter (BPF) between 800 and 2400 Hz and a high-pass filter (HPF) with cut-off frequency at 2400 Hz. For speech signals, those bands were chosen to manipulate three relevant frequency regions related to the fundamental frequency and second formant of voiced segments, and to the range of unvoiced fricatives, respectively. The use of the Gammatone filter bank provides more freedom to follow the spectral properties of the input (instrument) sounds. Another difference is that in our implementation we omitted the band level compensation (that would have come after Stage 3), which due to the large number of auditory bands in our algorithm (31 bands), introduced an increasing spectral tilt. The spectral tilt introduced a gradual increased band weighting towards the high frequencies with a relative emphasis of

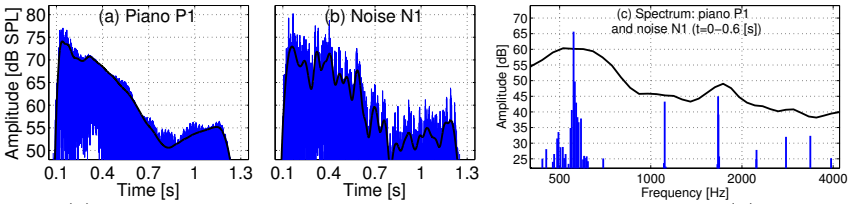


Figure 3.2: (a) Waveform of the Viennese piano P1 converted to SPL, and (b) one realisation of its resulting ICRA noise at an SNR= 0 dB. The thick black lines correspond to the Hilbert envelope of the waveforms (LPF with cut-off at 20 Hz). (c) Spectra of the piano sound (blue) and the ICRA noise (black thick line) averaged over the first 0.6 s of both waveforms.

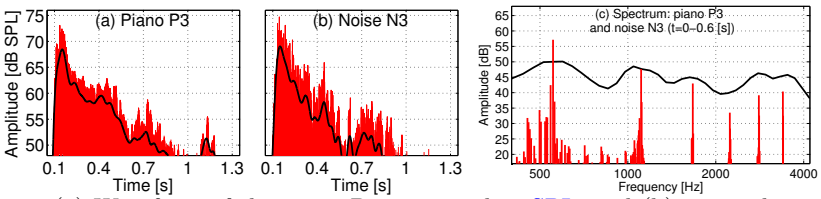


Figure 3.3: (a) Waveform of the piano P3 converted to SPL, and (b) one realisation of its resulting ICRA noise at an SNR= 0 dB. The thick black lines correspond to the Hilbert envelope of the waveforms. (c) Spectra of the piano sound (red) and the ICRA noise (black thick line) averaged over the first 0.6 s of both waveforms.

10 dB at the highest auditory filter with respect to the F0-centred band. This omission happened incidentally and we only became aware of it after the data collection. Some reflection about the spectral tilt is added in the discussion section and it is further investigated in Chapter 5.

3.2.2 Comparing two sounds

In this section we explain how the concept of ICRA noise can be used to compare two piano sounds. For this purpose, two recordings of the note C#$_5$ (nominal F0 of 554 Hz) from the pianos P1 and P3 were chosen (see Table 3.1). Firstly, the ICRA noise for both sounds has to be generated using the algorithm explained in the previous section. The resulting noises from the ICRA algorithm have an average (RMS) level that is the same as the level of the corresponding piano signals. At this level, the noises are interpreted to be at an SNR of 0 dB. The pianos P1 and P3 together with one realisation of their ICRA noises (N1 and N3) are shown in Figures 3.2 and 3.3. Since the sounds are compared pairwise, there are a number of considerations that have to be taken into account before conducting the experiment.

Practical considerations

During the experimental procedure, the task is to distinguish between two sounds. A three-alternative forced-choice (3-AFC) procedure is used.

This procedure is also known as odd-ball paradigm. In this procedure, one of the two sounds serves as "reference" and is presented in two observation intervals. The other sound serves as "test sound" and is presented in the randomly chosen third interval.

The sounds being compared need to be of a similar duration. In the example, both piano waveforms were set to have a duration of 1.3 s. Additionally the piano onset (leading to the maximum sound pressure level) was set to occur at approximately the same time stamp ($t = 0.1$ s).

The next consideration is to generate a "paired" ICRA noise that accounts for the spectro-temporal properties of both piano sounds. The paired noise is generated by combining the two ICRA noises (mean of their waveforms). The resulting noise is labelled as having an SNR of 0 dB[3]. It is also assumed that the paired ICRA noise is efficient to gradually mask the properties of the test sounds when presented together (in the example, P1 or P3 plus the paired noise) within each trial interval as the noise level increases (and the SNR decreases). It is important, however, to use different realisations of the paired noise in every test interval. This is because the use of a single fixed noise removes the statistical variability of the masker and may introduce additional cues during the course of the experiment (von Klitzing & Kohlrausch, 1994). The use of a fixed noise is known as frozen noise. If additional decision cues are available to the participant, the discrimination of the pianos becomes easier. To avoid this problem, noises that are independently generated but being drawn from the same statistical distribution are used. Such type of noises are known as running noises. To generate "running" ICRA noises, twelve realisations of each paired ICRA noise were generated. Within each trial of the 3-AFC experiment three paired noises are chosen, which leads to "12 choose 3" or $\binom{12}{3} = 220$ possible triads of noises. If the selection of noises is randomly drawn from a uniform distribution, it is unlikely that two participants use exactly the same sequence of paired noises during the course of the experimental session. In order to perform the actual comparison between the pianos P1 and P3, the SNR of their paired ICRA noises is adapted by applying a positive gain (decrease of the SNR, more difficult discrimination) or a negative gain (increase of the SNR, easier discrimination), depending on the participant's responses.

[3]By averaging the two waveforms the variance of the resulting paired noise is decreased by 3 dB.

3.2.3 Adaptive procedure: Instrument-in-noise test

The instrument sounds are compared pairwise. A given pair of sounds is presented in 3-AFC trials, where the discriminability threshold is estimated by adjusting the noise level. This corresponds to an adaptive procedure (or staircase method). The participant has to indicate which of the three intervals contains the target sound (presented once) where the reference sound is presented twice. The adjustable parameter (noise level) is varied following a two-down one-up rule: the noise is increased (SNR is decreased) after 2 consecutive right answers and decreased (SNR is increased) after 1 wrong answer. This paradigm tracks the 70.7% discriminability threshold (Levitt, 1971). Consecutive changes of the adaptive parameter in only one direction are "one run". A down run represents consecutive changes of the noise towards more difficult conditions (decrease in SNR) while an up run is related to consecutive changes towards easier conditions (increase in SNR). Changes from down to up (correct to incorrect) or up to down (incorrect to correct), the reversals, are the relevant noise conditions used as criterion to stop the experimental procedure. We chose to wait until 12 reversals are reached before stopping the comparison between the test sounds. The starting point of the paired ICRA noise is set to an SNR of 16 dB. We assume that at this SNR the discrimination of most piano pairs is easy and that this can help participants to get somehow accustomed to differences between the pianos being tested. The step size at which the noise is adjusted is set to 4 dB and is reduced to 2 dB (after the 2$^{\text{nd}}$ reversal) and 1 dB (after the 4$^{\text{th}}$ reversal). After the 4$^{\text{th}}$ reversal the runs stay at a fixed step size of 1 dB. These runs correspond to the measuring stage. The median of the reversals during the measuring stage (last 8 reversals) is used to estimate the discrimination threshold.

The sounds used in this chapter differ considerably in their loudness due to differences in the construction of the pianos from where they were recorded, which was affected by the fast technological developments during the 19$^{\text{th}}$ century. Loudness cues[4] are, however, not the main focus of this research. To avoid the use of loudness cues during the

[4]Three technical aspects that influence the loudness of the piano sounds are: (1) Differences in the force with which the hammer strikes the strings. One of the reasons for these differences is the use of different types of hammer actions, as it is the case for pianos P5 and P6 (see Table 3.1); (2) Differences in the radiation pattern of the pianos. This is influenced by the soundboard design, which differs from piano to piano; (3) Differences in string-soundboard coupling. This can be due to differences in both string and soundboard impedances at their coupling point. These three aspects do not only introduce loudness cues but also timbre (colour) cues. This means that despite the reduction of loudness cues in the experimental design, these aspects are at least partly present in the sounds due to their influence on timbre.

experiment, the stimuli were loudness balanced and the presentation level of each interval (piano + noise) was randomly varied (roved) by levels in the range ±4 dB, drawn from a uniform distribution. Additionally, explicit instructions were provided to the participants to not use level as discrimination criterion. The intervals lasted 1.3 s with an interstimulus interval of 0.2 s. During the course of the pilot experiments, an average answer period of 6 s was obtained. Therefore, every trial was expected to have a duration of about 11 s. The number of trials per comparison and per subject was variable and it was estimated to have an average of 45 trials per staircase. The evaluation of one pair of sounds takes about 8 minutes. This means that the method requires a long testing time to compare all the possible pair combinations within the dataset. With a dataset of 7 sounds, the number of pairwise comparisons (with no permutations) is $\binom{7}{2} = 21$, requiring almost 3 hours per participant to test the whole dataset. A balanced subset of data is considered to reduce the experiment duration. This is detailed later, in Section 3.3.

3.2.4 Reference procedure: Method of triadic comparisons

The method of triadic comparisons provides a way to obtain similarity judgements between elements without the need of verbal scaling techniques or actual physical measurements on the stimuli (Levelt et al., 1966; Shepard, 1987). The method has been used to successfully represent both perceptual and cognitive information in different research fields (see, e.g., Shepard, 1987; Burton & Nerlove, 1976). The method of triadic comparisons is, therefore, a well accepted method to evaluate similarity that has also been used in the assessment of perceptual spaces using sound stimuli (Levelt et al., 1966; van Veen & Houtgast, 1983; Fritz et al., 2010; Novello et al., 2011). For the previous reasons we chose this experimental procedure as a reference to validate the suggested instrument-in-noise method.

In the method of triadic comparisons, each trial consists of three sounds, namely, "A", "B", and "C". From this triad, three pairs can be formed: AB, AC, and BC. The task of the participant is to indicate which of the three pairs contains the most similar sounds and which one contains the least similar sounds. The remaining pair is labelled as having intermediate similarity. The participant can freely listen to each sample as many times as he or she needs. By presenting all the possible triads within a dataset, the participant's responses can be summarised in a similarity matrix. With a dataset of 7 sounds, the number of possible

triads is $\binom{7}{3} = 35$. Within the 35 triads, each of the 21 possible piano pairs is judged 5 times. The average time required to judge each trial, i.e., one triad, was 40 s meaning that a duration of about 23 minutes was expected to evaluate the whole dataset once.

One method to further process the experimentally obtained similarity matrix is the MDS algorithm (Shepard, 1962; Kruskal, 1964a, 1964b). MDS is commonly used as a visualisation tool of complex data. The similarity matrix is an $n \times n$ matrix (7×7 if $n = 7$ elements). In the MDS algorithm, the similarity matrix is assigned to a lower-dimensional space ($n \times q$ matrix), where the distance between elements is related to the perceptual similarity between them. The Euclidean distance between two elements in the q-dimensional space is a reference for the discrimination threshold estimated in the instrument-in-noise test.

3.3 Study case: Similarity among 19$^{\text{th}}$-century Viennese pianos

3.3.1 Stimuli

Recordings from seven pianos are compared among each other. The pianos were constructed in Vienna between 1805 and 1873. During this historical period, the piano construction underwent major developments. One important change during the 19$^{\text{th}}$ century was the increase of the string tension at rest (by a factor of 4), with the purpose of increasing the sound power of the piano. The soundboard, responsible for the sound radiation to the air, increased in thickness to withstand the higher string tensions together with the inclusion of metallic parts after 1850. The excitation mechanism of the strings (the hammer) increased systematically its mass to increase the amplitude of the hammer impact (Chaigne et al., 2016; Chaigne, 2016). These changes affected the timbre (or colour) of the radiated piano sounds. We believe that these seven pianos are a representative sample of the timbre changes of the instrument.

Recordings of one note (C#$_5$, F0 of 554 Hz) from the seven pianos were used. One recording per piano was chosen leading to a total of 7 stimuli. The duration of each waveform was set to 1.3 s, with the note onset occurring at a time stamp of 0.1 s. The sounds were ramped down using a 150-ms cosine ramp. The loudness of the sounds was adjusted to have a maximum value of 18 sone. For that purpose the short-term loudness from the time-varying loudness (TVL) model (Glasberg & Moore, 2002)

Table 3.1: List of pianos used in the listening experiments. Information about the intensity of the sounds is shown. The loudness of the sounds when presented 4 dB softer and 4 dB harder are shown in parentheses.

		Level [dB SPL]	Loudness [sone]
ID / Year	Manufacturer	L_{max} / L_{eq}	S_{max} / S_{avg}
P1 / 1805*	Gert Hecher	77.2 / 62.8	17.4 (13.7-22.0) / 6.8 (5.2-8.8)
P2 / 1819	Nannette Streicher	74.9 / 58.8	17.2 (13.5-21.8) / 5.5 (4.2-7.2)
P3 / 1828	Conrad Graf	73.7 / 55.4	17.0 (13.3-21.5) / 5.6 (4.3-7.3)
P4 / 1836	Johann B. Streicher	83.7 / 66.3	18.5 (14.4-23.5) / 7.0 (5.3-9.1)
P5 / 1851**	Johann B. Streicher (English)	78.0 / 60.2	17.8 (14.1-22.4) / 6.6 (5.1-8.5)
P6 / 1851**	Johann B. Streicher (Viennese)	81.7 / 67.2	17.2 (13.5-21.8) / 7.3 (5.6-9.1)
P7 / 1873	Johann B. Streicher & Sohn	81.7 / 67.2	17.4 (13.7-22.1) / 8.3 (6.3-10.7)

(*) Piano P1 is a contemporary replica of a piano built in 1805. (**) Pianos P5 and P6 differ in their hammer action (English and Viennese, respectively).

was used. After the adjustment, the sounds had a maximum level ranging from 73.7 to 83.7 dB SPL (see Table 3.1).

In order to compensate for pitch differences in the piano recordings, the mean pitch of the sounds was adjusted to 554 Hz. The maximum pitch difference was for pianos P3 and P7 which had a mean pitch of 519 Hz and no pitch adjustment was needed for the recording of Piano P6. The pitch adjustment was performed for each piano sound in two steps. In step one, the pitch of the sound was scaled to the desired value by using resampling. In step 2, a time stretch technique was used to keep the duration of the pitch-adjusted sounds constant. The time stretch was done by using the phase vocoder algorithm (Ellis, 2002)[5].

3.3.2 Apparatus

The experiments were conducted in a doubled-walled sound-proof booth. The stimuli were presented via Sennheiser HD 265 Linear circumaural headphones in a diotic reproduction (identical left and right channels). The participant's responses were collected on a computer using the software APEX (Francart et al., 2008) and the APE Toolbox for MATLAB (De Man & Reiss, 2014) for the instrument-in-noise and the triadic comparisons, respectively.

3.3.3 Participants

Twenty participants (8 females and 12 males) were recruited from the JF Schouten subject database of the TU/e university. At the time of testing, the participants were between 19 and 38 years old (average of 25) and they all had self-reported normal hearing. They provided their

[5]The phase vocoder algorithm is available at http://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc/ (Last accessed on 18/07/2018).

informed consent before starting the experimental session and were paid for their contribution.

The sample size of 20 participants was assessed a priori aiming at testing the hypothesis that the data from the instrument-in-noise are highly correlated (effect size or Pearson correlation of at least 0.6) with the data from the triadic comparisons, with a power of 90%. This analysis was done in the software G*Power (Faul et al., 2007, 2009), requiring 17 participants to reach the desired effect size. By increasing the number of participants to 20 the observable effect size is reduced to 0.57.

3.3.4 Experimental sessions

The experimental sessions were organised in two one-hour sessions per participant, including breaks. For the instrument-in-noise test, each participant was asked to evaluate 11 piano pairs. This means that the whole dataset (21 piano pairs) is tested once every two participants, including one common pair. For evaluating half of the dataset, a time of 1:30 hours was estimated. For the triadic comparisons a duration of 24 minutes was estimated. Participants were encouraged to take breaks if they felt tired or distracted, which may have resulted in longer and less accurate threshold estimations. The participants started the first session with the evaluation of 17 randomly chosen triads. This served as a way of familiarising the participants with the set of piano sounds. The session continued with 5 or 6 threshold estimations (staircase procedure) that always started at a low noise level (high SNR). Participants were not allowed to repeat the trials and no feedback was provided about the correctness of their responses. During the second session the participants evaluated the remaining 18 triads, followed by 6 or 5 threshold estimations, completing the total of 11 estimations. Two (or three) piano pairs were evaluated within the same experiment at a time, i.e., trials from 2 (or 3) staircases were interleaved. This means that the participant did not necessarily judge the same piano pair in consecutive trials. For choosing the distribution of piano pairs throughout the test, the order of the 21 pairs was randomised 5 times. Each randomisation was used to assign the piano combinations of 4 participants. Two participants tested the same piano pairs but exchanging the test and reference sounds. For instance if the piano pair 57 (piano P7 being the reference sound) was attributed to the one participant then the pair 75 (piano P5 being the reference sound) was attributed to the other participant. Two participants tested the first 11 pairs of the randomisation and two participants

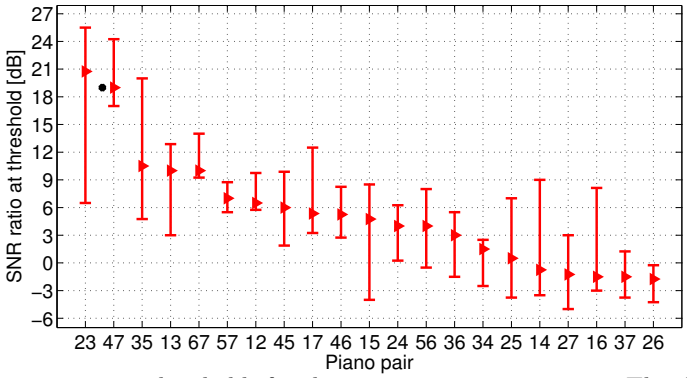


Figure 3.4: Discrimination thresholds for the instrument-in-noise tests. The thresholds (red triangles) are used as measure of similarity between the sounds and were assessed taking the median across participants. The piano pairs are shown along the abscissa and are ordered from higher to lower SNR thresholds. The error bars represent IQRs.
(*) The results for piano pair 47 consider 8 thresholds, with 3 estimations using the staircase procedure and 5 using a constant-stimulus procedure. See the text for further details.

tested the remaining 10 pairs of the randomisation plus one "common pair" (total of 11 pairs). With this distribution method and after finishing all the experimental sessions, each piano pair was tested 10 times with each piano sound in the pair being used 5 times as reference and 5 times as test sound. For the common pairs (5 in total), two additional comparisons were available, being evaluated 12 times. With this configuration, the whole dataset was tested 10 times including 5 pairs that were additionally tested twice. The expected number of estimations was therefore 220.

3.4 Results

3.4.1 Instrument-in-noise test

The discrimination thresholds of the instrument-in-noise experiment are shown in Figure 3.4. The pooled thresholds were assessed by taking the median of all individual threshold estimations per piano pair. No distinction was made between permuted piano pairs (e.g., pair 23 and pair 32 were pulled out together). The thresholds ranged from 20.75 dB for pair 23 down to -1.75 dB for pair 26. The estimations had a large between-subject variability with a length of the IQRs from 19.0 dB (pair 23) down to 3.25 dB (pair 57) with a median value of 8 dB. The results are based on 179 staircase threshold estimations and 5 threshold estimations using a constant stimulus procedure. During the data collection 210 of the 220 originally planned staircases were obtained. Ten thresholds were not estimated: for pair 47 five staircases were not conducted being

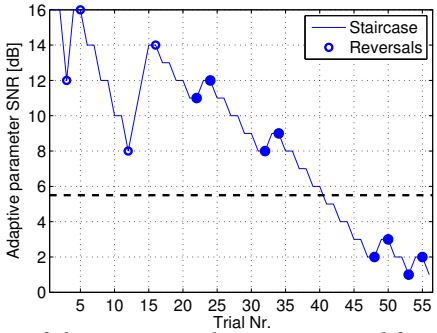


Figure 3.5: Example of one of the staircases that was removed from the data analysis. In this case, the last 4 reversals (SNRs at around 2 dB) differ in more than 3 dB from the estimated threshold (SNR at 5.5 dB), that considered the last 8 reversals (filled circle markers).

replaced by results obtained from a constant stimulus procedure at an SNR of 20 dB, while for participant S14 five piano pairs were accidentally skipped. For her, in session 1 and session 2 the same 6 pairs were tested. Only her results from session 1 were used in the data analysis. The results from session 2 were consistent and differed by no more than 2 dB with respect to the thresholds obtained in session 1. From the 210 obtained threshold 31 estimations were excluded.

Exclusion criteria

Thirty-one staircases were excluded from the data analysis after the data collection. Three staircases were incomplete, having less than 12 reversals. Three staircases were removed because the participants reached a maximum SNR of 50 dB ("minimum" noise level). This value was set in advance as floor condition. Participants reaching this point were not able at all to discriminate the two sounds being tested. The remaining 25 thresholds were removed after a check of consistency of the staircases. For this the standard deviation of the reversals was assessed. Thresholds estimations where the deviation of the reversals was larger than 3 dB were removed. The removed thresholds were checked manually to confirm that the staircase did indeed include inconsistencies between the convergence point of the staircase and the estimated threshold. Such a situation is illustrated in Figure 3.5 where one of those staircases is shown. This staircase has a convergence point (see the last four reversals) that differs from the threshold estimation by 3.5 dB.

Thresholds using a constant stimulus procedure

The evaluation of piano pair 47 (and 74) was for several participants very difficult. As part of our hypotheses the discrimination of sounds at high

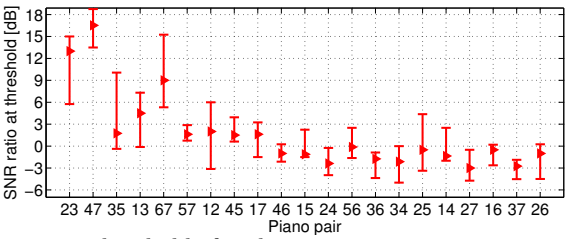


Figure 3.6: Discrimination thresholds for the instrument-in-noise test after applying a correction to account for the participant's variability. The thresholds (red triangles) are sorted as in Figure 3.4. The median length of the IQRs (across pairs) is 4.5 dB.

SNRs should be easy, with scores of nearly 100%. This was not the case for pair 47, where two staircases obtained from the first five participants had to be excluded according to the criteria described above. The level of the noise during the discrimination task was, on average, at levels around or above an SNR of 20 dB. This means that at an SNR of 20 dB, where we expected nearly perfect performance, the scores were often lower than the target score of 70.7%. For this reason, we decided to implement a constant stimulus experiment, where sixteen 3-AFC trials of pair 47 (or 74) were presented at an SNR of 20 dB. The percentage score could give an indication about how far away from that noise level the discrimination threshold could be expected. The scores obtained for the remaining 5 participants were 81.25, 50, 81.25, 50 and 68.75%. We were able to test pair 47 using the constant stimulus procedure at 20 dB with one participant of the first group (participant S06). The participant had an estimated adaptive threshold at 23.5 dB and the score obtained at 20 dB was 56.25%. This means that the participant's performance improved from 56.25% at 20 dB to 70.7% at 23.5 dB, which represents an average score increment of 4.1%/dB. This rate can be interpreted as the slope of the individual psychometric function for participant S06. We assumed, however, that this slope is also valid for other participants. In this way we converted the constant-stimulus scores of 81.25, 50, 81.25, 50, and 68.75% into the SNR thresholds of 17.5, 25.0, 17.5, 25.0 and 20.5 dB, respectively. These results were added to the raw thresholds results from the staircases. In spite of the lack of experimental evidence for this assumption, simulated thresholds (as in Chapter 4) showed that for piano pair 47, the scores increased at a similar rate of 4.6% (increase from 51.4% at 15 dB to 74.3% at 20 dB).

Between-subject variability

In order to understand the observed variability in the results of Figure 3.4, we first assessed the median of the estimated thresholds per

participant. Since the difficulty in the judgement of the piano pairs should be distributed across the 11 pairs evaluated per participant, the median thresholds give an indication about how sensitive each participant was during the course of the experiments. A lower median threshold indicates more sensitivity and, correspondingly, a higher threshold indicates less sensitivity to the cues available in the piano waveforms. The lowest and highest median SNR thresholds were found for participants S14 (avg. SNR=−7.25 dB) and S10 (avg. SNR=18.5 dB), with a median SNR across participants of 4.0 dB. This supports the existence of a strong difference in the participant's sensitivity. The SNR thresholds after a correction factor is applied are shown in Figure 3.6. The correction depends on the median participant thresholds. For instance, for the thresholds of participants S14 and S10 a correction of +7.25 and −18.5 dB (additive inverse values) was applied. With the correction the median length of the IQRs decreased from 8.0 to 4.5 dB. Although several piano pairs changed their rank order (the thresholds in Figure 3.6 are not monotonically decreasing), the rank-order correlation indicate a strong relationship of $r_s(19) = 0.83$, $p < 0.001$[6], between the thresholds before and after being corrected. This small effect is caused because the correction moved the pairs around but only in neighbouring relative locations. Despite the fact that with the results shown in Figure 3.6 the between-subject variability is almost halved, they are not used for any further processing in this chapter. We assume that the choice of the median as measure of central tendency of the thresholds is robust enough to deal with the large IQRs and that the results without correction (Figure 3.4) are representative.

3.4.2 Triadic comparison

The results of all participants were pulled out to construct the similarity matrix shown in the upper right triangle of Table 3.2. All participants judged the whole dataset of 35 possible triads once. Within the 35 triads the 21 pairs were judged 5 times. These numbers are relevant to understand the range of possible scores in the similarity matrix.

Construction of the similarity matrix

The similarity matrix is a way to summarise how often each piano pair was chosen as most similar, most dissimilar or indirectly chosen as having an intermediate similarity, when presented in triads with the other test

[6]The value between brackets indicate the degrees of freedom, which is $N - 2$, with N being the number of data points being compared.

Table 3.2: The similarity matrix S_{ij} derived from the responses of 20 participants (S01-S20) is shown in the upper right triangle. The maximum possible score is 200. The lower left triangle corresponds to the Euclidean distances between stimuli in the resulting four-dimensional space. A high score in the similarity matrix should correspond to a short Euclidean distance. The lowest and highest scores were obtained for the pairs 24 ($S_{ij} = 33$) and 23 ($S_{ij} = 190$). The corresponding distances were 0.91 and 0.26, respectively. The shortest distance was found for pair 47 ($S_{ij} = 189$) with a value of 0.14.

	Piano						
Piano	P1	P2	P3	P4	P5	P6	P7
P1	-	88	123	76	95	149	100
P2	0.75	-	190	33	79	54	45
P3	0.63	0.26	-	52	116	63	58
P4	0.78	0.91	0.86	-	119	103	189
P5	0.72	0.78	0.66	0.63	-	137	110
P6	0.51	0.86	0.83	0.69	0.56	-	121
P7	0.70	0.88	0.84	0.14	0.67	0.62	-

pianos. To score the results of each triad, 2 points were attributed to the pair indicated as most similar, no points to the least similar pair, and 1 point to the remaining pair. Since each pair of piano sounds was tested 5 times by 20 participants, the maximum possible score of a given pair is $S_{\max} = 200$ ($5 \times 20 \times 2$). The similarity matrices in the studies by Levelt et al. (1966) Fritz et al. (2010), Novello et al. (2011), and van Veen and Houtgast (1983) were constructed in a similar way.

Multidimensional scaling

To further process the experimental data, the similarity matrix was first converted into a measure of dissimilarity by using:

$$D_{ij} = \sqrt{1 - S_{ij}/S_{\max}} \tag{3.1}$$

with S_{ij} being each element of the similarity matrix, $S_{\max} = 200$ being the maximum possible score (for 20 participants), and D_{ij} being the elements of the new dissimilarity matrix. The dissimilarity matrix was then used as input for the classical (non-metric[7]) MDS algorithm available in the MATLAB Statistics toolbox. In the classical MDS algorithm the search of the reduced space with q dimensions (with $q < n = 7$), the eigenvectors ($n \times n$ matrix) and eigenvalues λ_i ($n \times 1$ matrix) corresponding to the dissimilarities scores D_{ij} are calculated and then the q eigenvectors corresponding to the largest q eigenvalues are taken. Here

[7]The term "non-metric" refers to the fact that the MDS algorithm takes data that are non-metric, in our case similarity/dissimilarity data, while the resulting geometrical configuration represents a metric solution to fit the input data (Kruskal, 1964a).

we report two criteria to test the adequacy of a q-dimensional representation. The first criterion corresponds to the regular goodness-of-fit indicator in the classical MDS algorithm and is given by Equation 3.2. A value P_q of at least 80% is considered to produce an adequate fit of the data in the q-dimensional space (Everitt, 2005).

$$P_q = 100 \cdot \frac{\sum_{i=1}^{q} |\lambda_i|}{\sum_{i=1}^{n} |\lambda_i|} \tag{3.2}$$

The second criterion assesses a stress value S_t, which is obtained from a residual sum of squares between the dissimilarities D_{ij} and the Euclidean distances d_{ij} of the resulting q-dimensional space (Kruskal, 1964b). This is the goodness-of-fit measure that is typically used when applying other MDS algorithms and is given by Equation 3.3.

$$S_t = 100 \cdot \sqrt{\frac{\sum_{i<j} (D_{ij} - d_{ij})^2}{\sum_{i<j} D_{ij}^2}} \tag{3.3}$$

For different S_t-values there are accepted benchmarks of the goodness of fit: poor ($S_t = 20\%$), fair ($S_t = 10\%$), good ($S_t = 5\%$), excellent ($S_t = 2.5\%$), and perfect ($S_t = 0\%$). A perfect configuration means that the distances d_{ij} and the dissimilarities D_{ij} have a perfect monotone relationship.

When applying the classical MDS algorithm to the obtained dissimilarity matrix, the resulting space has $q = 4$ dimensions, with a total goodness of fit $P_q = 99.5\%$ and individual contributions per dimension of 53.5, 25.6, 14.3 and 6.1%. The four dimensional space has a stress $S_t = 3.1\%$ (close to "excellent"), with cumulative stresses of 21.9% for the first two dimensions ("poor") and 7.5% for the first three dimensions (between "fair" and "good"). The Euclidean distances of the fitted four-dimensional space are shown in the lower left triangle of Table 3.2. For ease of visualisation, only the first two dimensions ($P_{q,\text{cum}} = 79.1\%$, $S_t = 21.9\%$) of the fitted perceptual space are shown in Figure 3.7. Although this reduced representation provides a poor fit ($P_{q,\text{cum}} < 80\%$; $S_t > 20\%$), the overall distribution of the piano sounds in the four-dimensional space is not changed. There is a change, however, in the relative distances between points.

The Euclidean distances between pianos in the four-dimensional space are shown in the lower left triangle of Table 3.2 and they are indicated as filled square markers in Figure 3.8. The Euclidean distances range

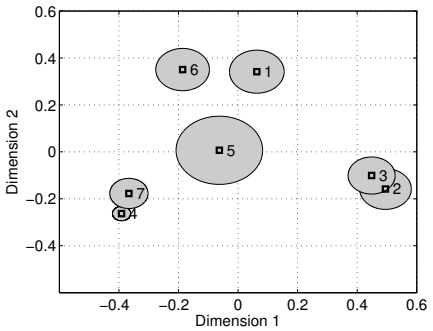


Figure 3.7: Perceptual space obtained with the classical MDS algorithm. Only the first two (out of four) dimensions are shown. This space suggests that the piano sounds (note C#$_5$) can be classified into four groups: pianos 16, 23, 47, and piano P5. Although the goodness of fit of this reduced representation is poor ($P_{q,\text{cum}} = 79.1\%$; $S_t = 21.9\%$) the overall distribution of the pianos in the space is not changed in the four dimensional space. The grey bubbles give an indication of the participant's variability: the bigger the bubble the higher the variability across participants. Note that the axes of the MDS space are not to scale.

between 0.14 (for pair 47) and 0.91 (for pair 24) with approximately 50% of the distances lying in the range between $d_{ij,25}$ =0.63 and $d_{ij,75}$ =0.83.

The results shown in Figure 3.7 suggest that the pianos (so far, limited to the note C#$_5$) can be classified into four distinct groups: pianos P1+P6, pianos P2+P3, pianos P4+P7 and piano P5. Although piano P5 seems to have an intermediate similarity with all these groups, in the four-dimensional space its distances increase systematically. The distances for all the other pianos do not differ considerably with respect to the ones in the two-dimensional representation.

Between-subject variability

The classical MDS algorithm does not provide any indication of the variability across participants in the resulting fitted space. One solution to this problem is provided by the individual differences scaling algorithm (INDSCAL) (Carroll & Chang, 1970). Within INDSCAL an individual perceptual space is assessed for every participant. Those spaces are assumed to be a weighted version of the resulting perceptual space, with different weights for different participants. With this approach it is possible to assess the stress of each stimulus per participant, which can be used for obtaining measures of variability. Although the data were processed using INDSCAL as implemented by de Leeuw and Mair (2009), this algorithm was finally not used because the fitted pooled space violated the condition of monotonicity between the D_{ij} and d_{ij}. An alternative approach was used that follows a similar idea to operate with the

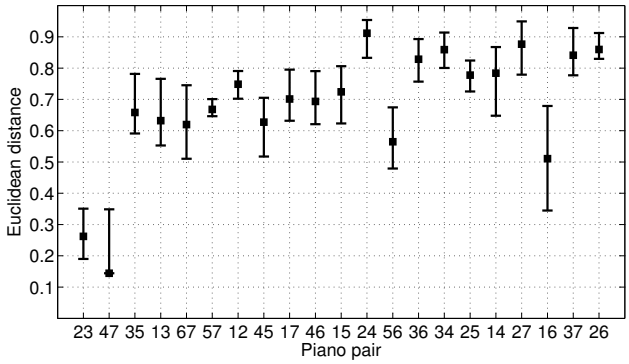


Figure 3.8: Euclidean distances taken from the four-dimensional perceptual space. These distances are also shown in the lower left triangle of Table 3.2. The piano pairs are sorted in the same way as in Figure 3.4. For a perfect consistency between these Euclidean distances and the instrument-in-noise results, the distances should increase monotonically. This does not happen but the correlation between distances and SNR thresholds are moderate to high, with values of -0.47 (Pearson) and -0.64 (Spearman) (see Figure 3.9). The error bars indicate the minimum and maximum distances between piano pairs across the 5 four-dimensional spaces assessed with data subsets every 4 participants.

stresses. Having as reference the fitted four-dimensional space, 5 dissimilarity matrices were generated pulling out the data of the participants S01-S04, S05-S08, S09-S12, S13-S16, and S17-S20, respectively. The classical MDS algorithm was applied, obtaining 5 new coordinates for each of the 7 test pianos. For each of the 7 pianos, the distances between these 5 coordinates and the coordinates in the pooled four-dimensional space was obtained, storing the difference between the minimum and maximum distances. Half of that difference is used as radius of the "bubbles" in Figure 3.7. The diameter of the bubbles has a median of 0.15, ranging from 0.06 (piano P4) to 0.29 (piano P5), which can be interpreted as piano P4 being judged more consistently across participants and piano P5 being scored more differently, leading to a higher between-subject variability. The obtained 5 four-dimensional spaces were used to assess the minimum and maximum distances between piano pairs and they are shown as error bars in Figure 3.8. Those deviations range between 0.05 (pair 57) and 0.33 (pair 16), with a median length of 0.17.

3.5 Discussion

A high perceptual similarity is equivalent to a high SNR threshold and a short Euclidean distance. If the results of both methods are consistent, the SNR thresholds of Figure 3.4, that are sorted in decreasing order, should correspond to monotonically increasing Euclidean dis-

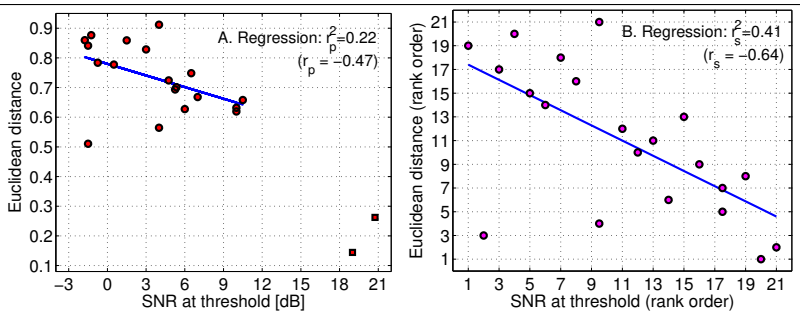


Figure 3.9: Regression between the instrument-in-noise and the triadic comparisons results. In panel A the scatter plot between SNR thresholds and Euclidean distances is shown. The results are significantly correlated with a Pearson $r_p(17) = -0.47$, $p = 0.04^*$. In panel B, the scatter plot of the rank order for the same data is shown. The results are significantly correlated with a Spearman $r_s(19) = -0.64$, $p = 0.001$. $^{(*)}$ The data of two pairs (pairs 23 and 47, panel A, square markers with $\text{thres}_{\text{exp}} > 18$ dB and $d_{ij} < 0.3$) had to be omitted to meet the normality assumption of the input data to the linear regression analysis (Pearson's correlation).

tances. Therefore, a perfect consistency between methods should be reflected by a correlation value of -1. Although the Euclidean distances shown in Figure 3.8 are not strictly monotonically increasing, the results have a significant moderate to high (Pearson) correlation of $r_p(17) = -0.47$, $p = 0.04$, and a high rank-order (Spearman) correlation of $r_s(19) = -0.64$, $p = 0.001$. The Pearson correlation tests whether the data are linearly related. Although this is an aspect that can be relevant, it imposes the assumption of normality on the data. To fulfil that assumption the data of two pairs (23 and 47) had to be omitted from the regression analysis. Since our data collection was designed to test an observable effect size of -0.57 (with 20 participants, see Section 3.3.3), the obtained r_p does not provide conclusive information about the relationship between SNR thresholds and Euclidean distances. For this reason, the Spearman correlation is more relevant because it does not require normally distributed data and it actually answers the question whether the assessed order of the samples (least to most similar or vice versa) is similar in both methods. In Figure 3.9, the SNR thresholds are shown on the abscissa and the Euclidean distances on the ordinate and they show the expected inverse relationship. The advantage of the Spearman over the Pearson correlation is reflected by the better distribution of the data along both the abscissa and ordinate axes in panel B of the figure.

Further inspection of the data shown in Figures 3.4 and 3.8 reveals that the two most similar pairs are the same in both methods (pairs 23 and 47). Both methods coincide in the judgement of 3 of the 6 most different pairs (thresholds < 0.5 dB and distances > 0.8): 26, 27, 37.

Piano P5 has an intermediate similarity with all the other pianos, with Euclidean distances between 0.56 (pair 56) and 0.78 (pair 15), this means that 5 (out of 6) distances are within the IQR of the distance data ($d_{ij,25-75} = 0.63 - 0.83$). This is also supported by the results of the instrument-in-noise test, where 5 (out of 6) thresholds lie within the IQR ($SNR_{25-75} = 0.2 - 7.7$ dB). For two pairs (16 and 56), both methods provide very different similarity measures. In both cases, the pairs are judged as being more similar in the triadic comparisons.

Although we hypothesised that the ICRA noises follow the spectro-temporal properties of the input piano sounds, as pointed out in Section 3.2.1, our algorithm "version A" introduced an incidental spectral mismatch that is gradual towards high frequencies. The effect of this spectral tilt is investigated in Chapter 5 and is compared with an updated version of the ICRA algorithm, "version B".

3.6 Conclusion

In this chapter we have presented a method to conduct a within-instrument comparison, measuring the perceptual similarity among test sounds using an instrument-in-noise test. In this method, the noise is matched to the spectro-temporal properties of the pair of sounds being tested.

Similarity among 19th-century Viennese pianos

As a study case, a comparison among recordings of one note ($C\#_5$) played on Viennese pianos from the 19th century was shown. The results of the instrument-in-noise test were compared with the results of the method of triadic comparisons, which is a method commonly used to map a set of stimuli into a perceptual similarity space. The results of both methods, collected from 20 participants, had a high and significant rank-order (Spearman) correlation of $r_s(19) = -0.64$, $p = 0.001$. The correlation results denote a high inverse relationship between SNR thresholds and Euclidean distances, meaning that a higher threshold results in a lower Euclidean distance. The results obtained from the instrument-in-noise method are consistent with overall subjective similarity judgements. Therefore, the instrument-in-noise procedure seems to be a promising method to quantify perceptual differences between sounds.

What is different when using the instrument-in-noise method?

It was pointed out that the instrument-in-noise method is rather time consuming when compared to the method of triadic comparisons (about

7 times slower), so why to choose it then? Despite the longer testing time, one of the advantages of the instrument-in-noise method is that it allows to measure similarity by evaluating different test conditions (different SNRs) where the physical properties of the test sounds are affected. This approach can be seen as a quantifiable way to manipulate the similarity between test sounds. On the contrary, the triadic comparisons are conducted at a fixed test condition (in our case in silence, i.e., at a very high SNR) and that leads (after data processing) to a purely psychological space where the physical properties of the sounds are kept constant. With this argument, the instrument-in-noise test can give an indication not only of which samples are closer or farther apart from each other (psychological approach), but can also provide evidence about their acoustic properties at noise levels below (SNRs above) and at threshold (physical approach).

Extending the use of the instrument-in-noise method

The key point of the instrument-in-noise method is the use of a noise that is shaped in spectral and temporal properties to the test sounds. The ICRA algorithm (Dreschler et al., 2001), used originally to generate speech maskers, was adapted to provide a suitable solution for instrument sounds. The described instrument-in-noise method can be used not only in the evaluation of other piano notes but also to evaluate any other instrument, as far as some practical aspects regarding the stimuli are followed. For the piano sounds, some of these aspects were: to have test stimuli with the same pitch, similar durations, a piano onset occurring at a "synchronised" time stamp, and to balance for any cue that is not desired to be judged (we kept the maximum loudness constant across stimuli). Some of the cues that were available to our participants were the envelope, attack and decay of the waveforms and their spectral content.

For the evaluation of other piano notes or other musical instruments, the ICRA noises have to be generated again in order to match the spectro-temporal properties of the "new" test sounds.

4 Simulating the perceived similarity of instrument sounds using an auditory model

In this chapter an auditory model which predicts psychoacoustic data is applied to the problem of perceptual similarity between complex sounds. The perceptual similarity task corresponds to the instrument-in-noise test presented and validated in Chapter 3. The same set of loudness-balanced piano sounds is used here.

The concept of similarity can be studied as a sensory process but, as argued in the next section, also as a cognitive process. The auditory model used in this chapter accounts primarily for the first aspect and also includes a "memory" stage that can be interpreted as a cognitive component within the model. The challenge of this chapter is the adjustment of the memory stage of the auditory model, i.e., the assessment and use of the so-called template of the system, in order to extend its use to account for the human performance in a similarity task using complex (piano) sounds.

4.1 Introduction

In the context of acoustics, similarity assessments are used in sound quality evaluation (see, e.g., Hansen & Kollmeier, 2000; Kates & Arehart, 2014) and in the study of specific sound types (see, e.g., Grey, 1977; Fritz et al., 2010). The study case of Chapter 2 is another example of this latter use. The concept of similarity is relevant because in an everyday listening experience, (sound) objects are unlikely to be repeated in exactly the same way (see, e.g., Shepard, 1987). Therefore, there is some acquired familiarisation used to recall those similar (sound) objects. For this reason, the concept of similarity has been studied as a cognitive or top-down process, reflecting the familiarisation with the object, as well as a perceptual or sensory process, reflecting how a given stimulus can "match" that object. In this chapter we use an auditory model

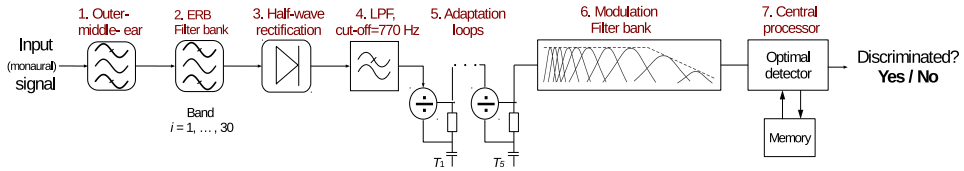


Figure 4.1: Block diagram of the PEMO model. Each of its stages is explained in the text.

that processes sounds primarily in a sensory fashion, but also includes a top-down (cognitive) component.

The auditory model used in this chapter belongs to the family of models of the "effective" processing of the auditory system. This set of models provides a unified framework to simulate a number of auditory phenomena such as simultaneous, backward, and forward-masking (Dau et al., 1996a, 1996b; Jepsen et al., 2008), modulation-detection (Dau et al., 1997a, 1997b; Jepsen et al., 2008), gap-detection (Münkner, 1993) and speech intelligibility by estimating speech reception thresholds (Dau et al., 1999; Ewert & Dau, 2000; Jørgensen & Dau, 2011). Unless otherwise specified, we will refer to this family of models as "auditory models" throughout this thesis. The specific auditory model that is used here is referred to as PEMO and it corresponds to the model described by Dau et al. (1997a) using the modulation filter bank set-up as described by Jepsen et al. (2008). The block diagram of the model is shown in Figure 4.1. We used the implementation of the PEMO model available within the AMT toolbox for MATLAB (Søndergaard & Majdak, 2013). In the AMT toolbox the peripheral stages of the model (stages 1-6 in Figure 4.1) are available. The peripheral stages deliver the internal representation of a sound. The last part of the model is an own implementation of the central processor. The central processor is a back-end stage that further compares two or more internal representations (obtained from two or more sounds processed within the PEMO model) with the aim of deciding whether those representations are distinct enough to be judged as "different" by a simulated human listener.

4.2 Description of the model

The input signal is a monaural sound with waveform amplitudes between -1 and 1[1]. Within the model an absolute amplitude of 1 (0 dBFS) is interpreted as a sound pressure level of 100 dB.

[1]The amplitude range between ±1 corresponds to amplitudes between ±32767 if the sounds are stored with an amplitude resolution of 16 bits ($2^{16} - 1 = 65535$ steps).

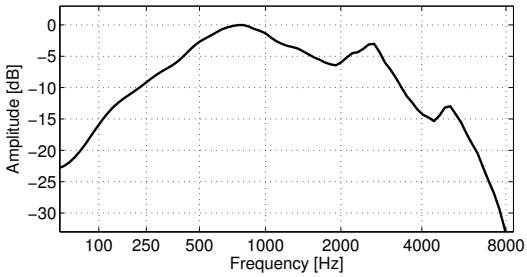


Figure 4.2: Combined frequency response of the outer- and middle-ear filters.

4.2.1 Outer- and middle-ear filtering

This stage accounts for the effects of the outer and middle ear on the incoming signal. The effects of both, the outer and middle ear, are implemented as 512-tap finite impulse response (FIR) filters. The outer-ear filter introduces a transfer function from headphones to the tympanic membrane, emphasising frequencies around 2750 Hz and attenuating frequencies above 6000 Hz (see Pralong & Carlile, 1996, their Figure 1(E)). The middle-ear filter introduces a transfer function from the tympanic membrane to the stapes. The output of this filter approximates the (peak-to-peak) velocity of the stapes in response to pure tones, that transfers oscillations into the inner ear through the oval window. This filter is based on Lopez-Poveda and Meddis (2001, their Figure 2) and Goode, Killion, Nakamura, and Nishihara (1994, their Figure 1, inset "Stapes (104 dB SPL)"). The combined response of the outer- and middle-ear filters is shown in Figure 4.2 and can be roughly described as a BPF centred at 800 Hz with slopes of 6 dB/octave below and above that frequency. This stage was also included in the auditory model of Jepsen et al. (2008) but not in previous versions of the PEMO model.

4.2.2 Gammatone filter bank

This set of filters corresponds to a linear approximation of a critical-band filter bank. The Gammatone filter bank consists of 31 bands having centre frequencies between 87 Hz (3 ERB_N[2]) and 7819 Hz (33 ERB_N), spaced at 1 ERB. The Gammatone filter bank is linear (it has a level-independent tuning). The PEMO model uses only the real part of the complex-valued all-pole implementation that is described by Hohmann (2002). All further processing stages of the model work independently on each auditory filter output.

[2]The ERB rate scale corresponds to one of the frequency scales that is inspired by the frequency representation in the auditory system. A brief overview of this scale is given in Appendix A.

4.2.3 Hair-cell transduction

This stage accounts for the inner hair-cell processing. It 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 low-pass filtered using 5 first-order infinite impulse response (IIR) filters with a cut-off frequency of 2000 Hz. The half-wave rectification keeps the positive part of the signal. The combined effect of the cascade of LPFs is equivalent to applying a fifth-order IIR filter with cut-off frequency of 770 Hz. With this LPF, the frequency components below 770 Hz are almost unaffected, so that the phase information is kept (a maximum attenuation of 3 dB is reached when approaching 770 Hz), frequency components between 770 Hz and 2000 Hz are gradually attenuated (attenuations between 3 dB at 770 Hz down to 15 dB at 2000 Hz), meaning that the phase information is gradually lost. For frequency components above 2000 Hz almost all the phase information is removed (more than 15 dB of attenuation, slope of -30 dB/octave). This way of removing phase information is consistent with the decrease of phase locking observed in the auditory nerve (Breebaart et al., 2001).

4.2.4 Adaptation

This stage simulates the adaptive properties of the auditory system at the level of the auditory nerve (see, e.g., Kohlrausch et al., 1992). Adaptation refers to changes in the gain of the system when the level of the input signal changes. When a change in the signal level is "rapid", the gain of the system remains constant and the level is transformed linearly. For slower variations, the signal level is compressed. This adaptation stage is implemented as 5 feedback loops, each of them having a different time constant ($\tau = 5, 50, 129, 253, 500$ ms). In this study an overshoot limitation is used, meaning that the output value for rapid input changes (relative to the time constants) is limited to a maximum value of 5 times the stationary output value for the same level. The limiter factor $\mathsf{limit}= 5$ differs with respect to the usual limiter factor of 10 used in the auditory models (Münkner, 1993; Dau et al., 1997a). Due to the relevance of the note onset in piano sounds, the choice of this new limiter factor is a sensitive parameter which strongly influenced the simulation results that are shown later in this chapter. The effect of using the new limiter factor on the resulting internal representations is described in the next section. The interested reader is also referred to Appendix C, where an in-depth review of the properties of the adaptation loops is given.

Table 4.1: Empirical parameters of the modulation filter bank. The cut-off frequencies f_{inf} and f_{sup} correspond to the −3 dB points of the transfer functions.

	Frequency [Hz]		BW			Frequency [Hz]		BW	
Nr.	mf_c	f_{inf} - f_{sup}	[Hz]	Q	Nr.	mf_c	f_{inf} - f_{sup}	[Hz]	Q
1	1.4	0.0 - 2.7	2.7	0.5	7	77.2	57.9 - 96.9	39.0	2.0
2	5.0	2.7 - 8.1	5.4	0.9	8	128.6	96.9 - 160.8	63.9	2.0
3	10.0	7.4 - 12.8	5.4	1.9	9	214.3	160.8 - 268.5	107.7	2.0
4	16.7	12.8 - 20.9	8.1	2.1	10	357.2	268.5 - 446.8	178.3	2.0
5	27.8	20.9 - 35.0	14.1	2.0	11	595.4	446.8 - 744.2	297.4	2.0
6	46.3	35.0 - 58.5	23.6	2.0	12	992.3	744.2 -1240.9	496.6	2.0

4.2.5 Modulation filter bank

The modulation filter bank corresponds to a linear filter bank that allows the processing of the incoming signal in terms of changes in its envelope. First, a reduction in the sensitivity to modulation frequencies above 150 Hz is introduced (Kohlrausch et al., 2000). For this purpose a first-order IIR filter with a cut-off frequency at 150 Hz and approximate roll-off of 6 dB/octave is applied. The filter bank comprises a maximum of 12 filters that have two different envelope frequency domains:

- Bands with modulation centre frequencies $\text{mf}_c \leq 10$ Hz (bands 1-3 in Table 4.1): the filters have a nominal bandwidth of 5 Hz (actual $BW = 5.4$ Hz). The first is an LPF with a nominal cut-off frequency of 2.5 Hz (actual $\text{mf}_{\text{cut-off}} = 2.7$ Hz). The real-valued part of the filtered signals is used, which corresponds to the band-limited output signal. This processing keeps the modulation phase information.

- Bands with modulation centre frequencies $\text{mf}_c > 10$ Hz (bands 4-12 in Table 4.1): the filters have a logarithmic scaling with a constant Q factor of 2 ($Q = \text{mf}_c/BW$). The absolute value of the complex output is used, which represents an approximation to the Hilbert envelope (Hohmann, 2002). This process reduces considerably the amount of modulation phase information but keeps the energy produced by the modulations within the respective band. An attenuation factor of $\sqrt{2}$ is applied to the resulting signals (Jepsen et al., 2008).

The modulation filters for each audio frequency band are limited to filters having an mf_c below a quarter of the audio centre frequency f_c. This is motivated by the results presented by Langner and Schreiner (1988), where evidence is provided that the neural activity in the auditory path (in the brain stem) has best modulation frequencies limited to that frequency range ($\text{mf}_c < f_c/4$).

4.2.6 Central processor

In this stage, the information received from the modulation filter bank is compared with a reference representation or "sound image" that is stored in the "memory" of the model. Inspired by the concept of an optimal detector, borrowed from signal detection theory (see, e.g., Green & Swets, 1966, their chapters 6 and 7), the model can be seen as an artificial listener[3] and the "memory" of the model can be seen as an expected sound representation, learned by experience, that gives a clear indication to the artificial listener about "what to listen for" (Green & Swets, 1966; Dau et al., 1996a). This memory is referred to as template.

In a 3-AFC task, there are three intervals that can be compared with the template. If the representations of each interval are labelled as R_x with $x = 1, 2, 3$, the interval having the highest similarity with the template would be always chosen by the artificial listener. One mathematical way to express this idea is to assess the cross-correlation value (CCV) between the representation R_x and the template T_p:

$$\mathrm{CCV}_x = \frac{1}{f_s} \sum_{n=1}^{N} R_x[n] \cdot T_p[n] \tag{4.1}$$

It is important to stress that the template T_p is a unit energy representation while the representation R_x is not. As explained in subsequent sections, however, a difference representation ΔR_x is used in this equation instead of the direct use of the representation R_x.

Memory: Use of a template

The use of a memory template, or simply template, assumes that in the detection of a signal (or object) among other signals (or objects), some type of awareness about the target signal is used. This corresponds to a top-down process and can be seen as a cognitive component in the auditory model. This approach is also used in the field of vision where there is evidence of brain activity in response to features of the expected signal (see, e.g., Chelazzi et al., 1993). The template is derived or, in other words, is "learned" by the artificial listener, at the beginning of the experiment simulation in a condition that is assumed to be easily detected (low-noise condition, high SNR). This condition is referred to as a suprathreshold SNR. In the simulations, the suprathreshold SNR was set to 21 dB. This condition is 5 dB higher in SNR (lower noise) than the initial SNR of the experimental sessions.

[3] In the literature (and in this thesis), the terms "artificial listener" and "artificial observer" are used interchangeably.

Template in a similarity task

The derivation of the template T_p in a similarity task where two sounds are compared is determined by: (a) the two test sounds, the target and "reference", and; (b) two or more realisations of a noise that can efficiently mask the properties of both piano sounds. To account for the latter aspect, ICRA noises (in this chapter using version A of the algorithm, as in Chapter 3) are used in every piano presentation. For the first aspect, the internal representations of both, the target piano $R_t(MT)$ and reference piano $R_r(MT)$ have to be used, because their discrimination threshold depends on how different they are from each other. The argument (MT) indicates that the derivation of T_p should be done at a highly discriminable noise level, that is, at a low noise level (high SNR). In the course of this research different ways of deriving the template T_p using $R_t(MT)$ and $R_r(MT)$ were evaluated. Except for the adopted approach which is described in this section, the alternative approaches are described in Appendix E.

In the adopted approach, two templates are derived: (a) $T_{p,t}$ for the target piano sound, and (b) $T_{p,r}$ for the reference piano sound. For each of the templates, an average representation of the piano sounds embedded in four different realisations of the ICRA noises at a highly discriminable condition (here at an SNR of 21 dB) is obtained[4]. The average representations are normalised to unit energy. In this way, the templates satisfy (see also Equation E.3, in Appendix E):

$$E_t = \frac{1}{f_s}\sum_{n=1}^{N} T_{p,t}^2[n] = 1$$
$$E_r = \frac{1}{f_s}\sum_{n=1}^{N} T_{p,r}^2[n] = 1 \tag{4.2}$$

where N corresponds to the number of samples used by the artificial listener to make the decision. If the artificial listener uses the whole piano waveforms, then N is defined by the total length of the sounds (1.3 s for the anechoic pianos)[5]. A longer "observation" (listening) period would mean that the listener makes use of the undershoot effect after

[4]The number of ICRA noise realisations (four) used to derive each average piano-plus-noise representation was an arbitrary choice.

[5]In analogy to the theory of optimal detectors presented by Green and Swets (1966), we treat the templates $T_{p,t}$ and $T_{p,r}$ as "expected signals" along one (temporal) dimension. In fact, there are two other dimensions: audio and modulation frequency. Considering all template dimensions and following the nomenclature of Equation 4.7, Equation 4.2 would turn into $E_t = \frac{1}{f_s}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{n=1}^{N} T_{p,t\ mk}^2[n] = 1$.

the piano sounds vanish. Based on our experimental design, where the piano intervals had an interstimulus time of 0.2 s, the maximum possible observation period is 1.5 s. The results in the subsequent sections show, however, that an actual observation period of 0.5 s or less provides a better fit between simulated and experimental thresholds than the use of full piano waveforms.

Use of two templates

In the course of the simulation of a 3-AFC task the "expected signals" or templates ($T_{p,t}$ and $T_{p,r}$) have to be compared with the intervals ($x = 1, 2, 3$) of each trial. The expression shown in Equation 4.1 is used for this purpose but using a difference representation ΔR_x instead of the direct use of the representation R_x. The representation ΔR_x is obtained as the difference between the "piano-plus-noise" representation R_x and the representation of the corresponding paired ICRA noise $R_{N,x}$ at the SNR of the ongoing trial, obtaining three ΔR_x representations[6].

Due to the use of two templates, six CCV values are obtained, with three CCV values corresponding to the comparison between each (difference) interval representation ΔR_x with the target template $T_{p,t}$ and three corresponding to the comparison with the reference template $T_{p,r}$:

$$\begin{aligned} \mathrm{CCV}_{x,t} &= \frac{1}{f_s} \sum_{n=1}^{N} \Delta R_x[n] \cdot T_{p,t}[n] \\ \mathrm{CCV}_{x,r} &= \frac{1}{f_s} \sum_{n=1}^{N} \Delta R_x[n] \cdot T_{p,r}[n] \qquad \text{with } x = 1, 2, 3 \end{aligned} \tag{4.3}$$

Based on these six CCV values, the artificial listener chooses the interval that is more likely to contain the target sound using two criteria. If we assume that the target interval is presented in the first observation interval, then for a correct discrimination:

$$\begin{aligned} \max\left\{\widehat{\mathrm{CCV}}_{x,t}\right\} &= \widehat{\mathrm{CCV}}_{1,t} \\ \min\left\{\widehat{\mathrm{CCV}}_{x,r}\right\} &= \widehat{\mathrm{CCV}}_{1,r} \qquad \text{with } x = 1, 2, 3 \end{aligned} \tag{4.4}$$

[6] The use of difference representations ΔR_x is relevant for our decision approach due to the use of two templates. Since the unit energy normalisation of the templates is done independently, the noise alone representations $R_{N,x}$ that are used in both criteria will always have different $\mathrm{CCV}_{x,t}$ and $\mathrm{CCV}_{x,r}$ values. Subtracting the noise alone representations in the CCV calculation implies that the resulting $\mathrm{CCV}_{x,t}$ and $\mathrm{CCV}_{x,r}$ values correspond to the contribution of information of piano x relative to the contribution of the noise x.

In other words, the target template $T_{p,t}$ is expected to elicit the maximum CCV value when being correlated with the *target* interval. Likewise, the reference template $T_{p,r}$ elicits higher CCV values when being correlated with the *reference* intervals and therefore the lowest CCV value is attributed to the target interval. The hat symbol indicates that the CCV values differ from the exact definition given in Equation 4.3. This is caused by an internal noise, whose values are drawn from a Gaussian distribution $N(\mu, \sigma^2)$ with mean $\mu = 0$ and standard deviation $\sigma = 10.1$ MU (variance of σ^2). In our implementation of the internal noise, three numbers are added to the corresponding CCV_x value:

$$\begin{aligned} \widehat{\text{CCV}}_{x,t} &= \text{CCV}_{x,t} + N_x(\mu, \sigma^2) \\ \widehat{\text{CCV}}_{x,r} &= \text{CCV}_{x,r} + N_x(\mu, \sigma^2) \qquad \text{with } x = 1, 2, 3 \end{aligned} \tag{4.5}$$

Since $\mu = 0$, the standard deviation σ corresponds to the actual source of internal variability in the decision process. The use of this Gaussian noise leads to a reduction in the process performance when either the $\text{CCV}_{x,t}$ values get close to each other or when the $\text{CCV}_{x,r}$ values do. The standard deviation $\sigma = 10.1$ Model Units (MU) was obtained by running an increment-discrimination task with each piano sound and tracking the amount of noise needed to produce an average performance of 70.7% for a difference in level of $\Delta L = 1$ dB. This procedure is described in Appendix D.

Compensating the misalignment between piano representations

One final aspect in the decision criterion is that the cross-correlation between the templates and the interval representations should deliver the highest CCV values. As described in Section 3.3.1, the piano stimuli are aligned to have the note onset at a time stamp of 0.1 s. This alignment criterion seemed to be enough to perceive each of the piano sounds aligned with the ICRA noise within each piano-plus-noise interval. However, this does not always ensure a maximum CCV value during the decision process. This is particularly sensitive when correlating either the target piano representation with the reference template $T_{p,r}$ or the reference piano representation with the target template $T_{p,t}$. During the simulations, the cross-correlation function is assessed for each interval, with time lags between -50 ms and 50 ms (in steps of 1 ms). The maximum of the cross-correlation function is used as the CCV_x value for the decision stage (Equation 4.5).

4.2.7 Sources of internal and external variability

The processing of sounds in the auditory system is influenced by uncertainties in the stimuli and by internal variability caused, e.g., by imperfections in memory and changes in the level of concentration (see, e.g., Yost et al., 1989). In this thesis we differentiate between sources of variability that are internal or external. Uncertainties in the stimuli are related to an external source of variability, while the effects of human memory and concentration correspond to sources of internal variability. To (partly) account for variations in human performance due to sources of internal variability, an internal noise is often used within computational frameworks of auditory processing. In our model implementation, the internal noise is simulated by adding a Gaussian noise $N(\mu, \sigma^2)$ with mean $\mu = 0$ and standard deviation $\sigma = 10.1$ MU (see Equation 4.5 and Appendix D). In threshold-detection tasks a typical source of external variability is the use of running noise. Running noise refers to the fact that in different intervals of a trial, different realisations of similarly generated noises are used. In the instrument-in-noise test a running noise condition is approximated by using 12 different ICRA noise realisations for each piano pair. Another source of external variability in the instrument-in-noise test is the presentation level of each interval, which is randomised (roved) by levels in the range ±4 dB.

4.3 Description of internal representations

4.3.1 General description of the representations

The internal representation of pianos P1 and P3 after the last stage of peripheral processing of the auditory model (stage 6, modulation filter bank) is shown in Figure 4.3. The analysis is shown for one audio frequency band (centred at $f_c = 11$ ERB_N or 520 Hz, closest band to F0= 554 Hz). The piano sounds start at $t = 0.1$ s and their onsets occur shortly thereafter. The onset of the lowest modulation filter (Nr. 1, $\text{mf}_c = 1.4$ Hz) occurs approximately at $t = 0.20$ s, for filter Nr. 2 at $t = 0.15$ s and for the rest of the filters between t =0.10 and t =0.11 s. In the figure, it can also be observed that after the piano onset, the amplitudes in the modulation filters of P3 (Nr. 2-8) present more variations in comparison with piano P1, especially for $t = 1.0 - 1.3$ s.

We next describe the effect of using a stronger limiter factor in the adaptation loop stage. For this analysis the initial part (first 0.25 s of the waveform) of one of the piano sounds (piano P1) is further described. This description is also valid for the other 6 piano sounds of the dataset.

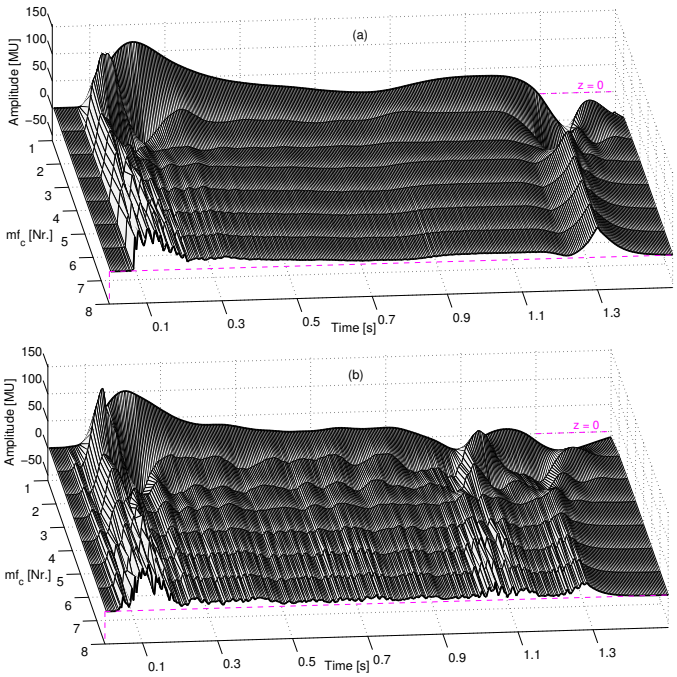


Figure 4.3: Internal representation for the recordings of piano P1 (panel (a)) and piano P3 (panel (b)). These internal representations correspond to the outputs of the peripheral stage of the PEMO model. For clarity, the analysis of only one of the 31 audio frequency bands (centred at $f_c = 520$ Hz, closest band to F0= 554 Hz) is shown. This band has 8 modulation filters with frequencies mf_c between 1.4 and 128.6 Hz).

4.3.2 Stronger limiter factor

The internal representation of the first 0.25 s of the piano P1 waveform is shown in Figure 4.4. The representation shown in panel (a) is a zoomed-in version of the representation shown in Figure 4.3(a). Two different configurations of the adaptation loops are used: using a limiter factor of 5 (as used in this thesis, panel (a) of Figure 4.4) and using a factor of 10 (as used in the literature, panel (b)). The representation considering the limiter factor limit= 5 (panel (a)) has amplitudes that range between −27 and 142 MU. The amplitudes of the representation with limit= 10 (panel (b)) range between −62.5 and 231.5 MU. In both cases the minimum and maximum amplitudes occur in the modulation filter Nr. 2, which is centred at $\mathrm{mf}_c = 5$ Hz. The difference between both representations is shown in panel (c) of Figure 4.4. Positive and negative amplitudes indicate that the less-compressed representation (limit= 10) has a wider range of amplitudes than those of the representation with limit= 5. The largest difference between the representations is found in the modulation

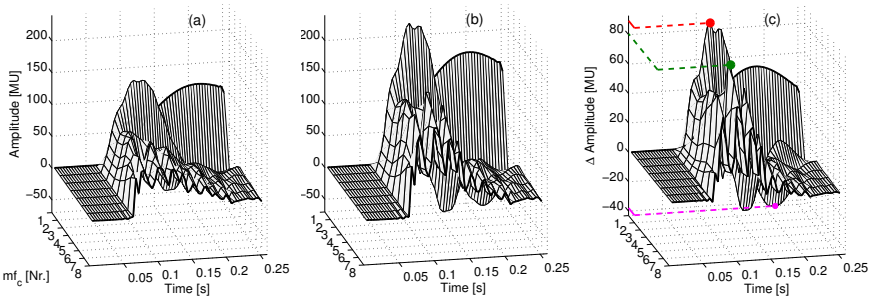


Figure 4.4: Internal representation of piano P1 (initial 0.25 s) with limiter factors of 5 (panel (a)) and 10 (panel (b)) in the adaptation loops stage. The 8 modulation filters that correspond to the audio frequency band centred at $f_c = 520$ Hz are shown. For both representations the minimum and maximum amplitudes are found for the modulation filter Nr. 2 (centred at mf_c of 5 Hz). The representation with limit= 5 has amplitudes that range between −27 and 142 MU. The representation with limit= 10 has amplitudes that range between −62.5 (not clearly visible) and 231.5 MU. In panel (c) the difference between both representations is shown. The maximum differences for low (89.5 MU, band Nr. 2) and high modulation bands (80.6 MU, band Nr. 6) are indicated by the red and green markers, respectively. The minimum difference of −37.9 MU is indicated by the magenta maker.

filter Nr. 2, where the representation with limit= 10 reaches an amplitude 89.5 MU above the maximum of the representation with limit= 5.

4.3.3 Information in the internal representations

In order to introduce an information-based analysis of the three-dimensional internal representations (dimensions n, m, and k), the following expression may be used:

$$I_m = 1/f_s \cdot \sum_{k=1}^{K} \sum_{n=1}^{N} R_{mk}^2[n], \qquad I_k = 1/f_s \cdot \sum_{m=1}^{M} \sum_{n=1}^{N} R_{mk}^2[n] \tag{4.6}$$

This expression is similar to Equation 4.3, but the subindexes m and k have been added to indicate that the sum, i.e., the "integration of information", can be done by either deriving the contribution (1) I_m of $M = 31$ audio frequency bands across all modulation filter bands, or (2) I_k of $K = 12$ modulation frequency bands across all audio frequency bands. In this section we express the contributions I_m and I_k as percentages of the total information I_{tot} available in the representation R:

$$I_{\mathrm{tot}} = \sum_{m=1}^{M} I_m = \sum_{k=1}^{K} I_k = 1/f_s \cdot \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{n=1}^{N} R_{mk}^2[n] \tag{4.7}$$

The results of this information-based analysis applied to the representation of piano P1 is shown in panels (a) and (b) of Figure 4.5 for the

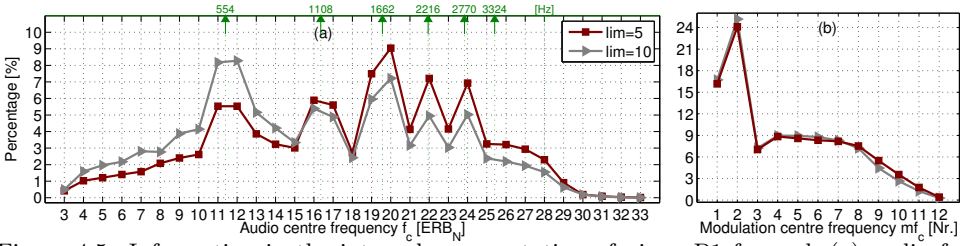


Figure 4.5: Information in the internal representation of piano P1 for each (a) audio frequency channel (I_m/I_{tot}), and (b) modulation frequency channel (I_k/I_{tot}). The maroon square markers indicate the information in the representation with limit= 5. The grey triangle markers indicate the information in the representation with limit= 10. The values per band are expressed as percentage with respect to the total information I_{tot}. The points along the ERB scale that correspond to F0= 554 Hz and its five first harmonics are indicated by the green labels on the top axis.

audio (I_m/I_{tot}) and modulation frequency bands (I_k/I_{tot}), respectively. It can be observed that the use of a stronger limiter factor of 5 increases the relative contribution of higher audio frequency bands, while no substantial change in the information weighting is observed across modulation filters. For the representation with limit= 5, the audio frequency bands with f_c below 15 ERB_N (924 Hz, containing the F0 of the piano note) comprise only 30.9% of the information in contrast to 45.6% for the representation with limit= 10 in the same frequency region. In terms of modulation frequency content, which is similar for both representations, bands 1 and 2 ($\text{mf}_c \leq 5$ Hz) comprise about 40% of the information and the remaining 60% is distributed across bands 3 to 12.

4.4 Comparison between experimental and simulated thresholds

4.4.1 Apparatus and procedure

The simulations were run using the AFC toolbox for MATLAB (Ewert, 2013). The AFC toolbox provides a framework to conduct listening experiments. The toolbox includes a feature where an artificial listener can be used during the experiments. The artificial listener uses an auditory model with a central processor based on signal detection theory. The PEMO model described earlier in this chapter was used.

The experiment was implemented as a 3-AFC task with the level of the ICRA noises used as adjustable parameter. The set-up of the task is similar to that used in the experimental sessions, which is described in Section 3.2.3. There are, however, small deviations from that descrip-

tion, mainly aiming at reducing the simulation time. Two sounds are compared: the target sound (presented once) and the reference sound (presented twice). The noise level was adjusted following a two-down one-up rule until 8 reversals are reached (4 reversals less than in the experimental sessions). The step sizes were set to 4 dB, 2 dB (after the second reversal) and 1 dB (after the fourth reversal). The median of the reversals during the measuring stage (last 4 reversals) is used to estimate the discrimination threshold of each pair of sounds. The presentation level of the sounds was randomly varied (roved) by levels in the range ±4 dB, drawn from a uniform distribution. The threshold estimation was repeated 6 times for each condition.

4.4.2 Stimuli

Piano sounds

The same selection of Viennese piano recordings as in Chapter 3 was used for the simulations. Recordings of the note C#$_5$ (F0= 554 Hz) from seven pianos were used. One recording per piano was chosen leading to a total of 7 stimuli. The sounds were set to have a duration of 1.3 s and they were ramped-down using a 150-ms cosine ramp. They had a maximum loudness S_{max} of about 18 sone (refer to Table 3.1). The pairwise comparison of all stimuli leads to a total of 21 possible combinations. For each piano pair 6 thresholds were simulated, 3 times the target piano was "A" and the reference piano was "B", the remaining 3 times the target piano was "B" and the reference piano was "A".

Piano-weighted noises

The same ICRA noises (version A[7]) generated for the listening experiments of Chapter 3 were used in the simulations. For the comparison of pianos "A" and "B" (or "B" and "A") individual noises that follow the spectro-temporal properties of each piano were combined to generate a paired noise AB (refer to Section 3.2.2 for further details).

4.4.3 Exploratory simulations: Subset of piano sounds

At first, a subset of 9 (of the 21) available piano pairs was used for the simulations. This selection was based on the results presented in

[7]In spite of the drawback in the spectral-matching properties (spectral tilt) of the ICRA-noise algorithm version A, identified and briefly introduced in Chapter 3, the same ICRA noises are used in this chapter. We assume that any effect of the spectral tilt on the experimental results of Chapter 3 should also be tracked in the simulations of the current chapter. An in-depth analysis of the spectral tilt effect is presented in Chapter 5 by means of simulations using noises with (version A) and without spectral tilt ("new" noises version B, adopted in Chapter 5).

Figures 3.4 and 3.8 of the previous chapter, from where 9 pairs that are well distributed along the abscissa, i.e., the "similarity" axis, were chosen. The selected piano pairs were: pair 12, 15, 16, 23, 26, 27, 37, 45, and 47. The pairs 23 and 47 were taken from the most similar end of the similarity axes. The pairs 26, 27, and 37 were taken from the least similar end of the axes. The remaining pairs 12, 15, 16, and 45 were taken from the intermediate similarity range.

This subset was used for (1) developing our template approach, and (2) testing the duration of the "observation (listening) period" of the template. This latter aspect is a consequence of the lack of success (see the last column of Table 4.2) to simulate the discrimination thresholds when using whole-duration piano waveforms as input to the model. The low thresholds in that condition were attributed to a sensitive artificial listener, who has access to more information than the human listeners. As a way to remove available cues within the auditory model, the piano sounds were truncated to shorter durations. This is equivalent to reducing the "observation" period t_{obs} of the artificial listener and can be seen as a simple way to account for a limited human-like working memory.

Under the hypothesis that participants provided a greater weighting to the note onset, a truncation of the piano waveforms would have to provide a higher correlation between the simulations and the experimental results. As will be shown in the results section, the simulation results provide evidence to support this hypothesis.

4.4.4 Simulations using the whole dataset of piano sounds

The simulation of discrimination thresholds $\text{thres}_{\text{sim}}$ for the whole dataset of piano sounds (21 piano pairs) was run using the optimal observation period t_{obs} obtained from the exploratory simulations and the adopted template approach. These $\text{thres}_{\text{sim}}$ values were used to evaluate the performance of the artificial listener with respect to the existing experimental thresholds $\text{thres}_{\text{exp}}$ (Chapter 3). In order to complement this evaluation, a comparison of $\text{thres}_{\text{sim}}$ values with Euclidean distances from two perceptual MDS spaces were also included: (1) from the space of Chapter 3, and (2) from a newly generated MDS space using the PEMO model. This newly generated MDS space is built by applying the described template approach to triadic comparison trials using the dataset of piano sounds.

Simulation of triadic comparisons using the current template approach
As described in Section 3.2.4, each triadic comparison trial consisted of three sounds that are labelled as "A", "B", and "C". From this triad, out of the three pairs that can be formed (AB, AC, BC) the participant (the artificial listener) has to indicate which of the three pairs contains the most similar sounds and which one contains the least similar sounds. In this way, the remaining pair is labelled as having intermediate similarity.

To simulate this task, the whole dataset of pianos ($\binom{7}{3} = 35$ triads) was used being restricted to the optimal t_{obs} duration. No noise was used because the experimental triadic comparisons were conducted in silence. Within each trial, three templates T_A, T_B, and T_C were derived by normalising to unit energy the corresponding internal piano representation R_A, R_B, and R_C. Two CCV values per pair were assessed (AB, AC, BC). For pair AB:

$$\text{CCV}_{AB} = \frac{1}{f_s} \sum_{n=1}^{N_{\text{obs}}} R_A[n] \cdot T_B[n], \qquad \text{CCV}_{BA} = \frac{1}{f_s} \sum_{n=1}^{N_{\text{obs}}} R_B[n] \cdot T_A[n]$$

where N_{obs} corresponds to the number of samples used by the artificial listener to make the decision and is related to the optimal observation duration t_{obs}. The CCV values for pair AC and BC can be obtained in a similar manner. Finally, three $\widehat{\text{CCV}}$ values were obtained, one for each pair:

$$\begin{aligned}
\widehat{\text{CCV}}_{AB} &= \max\{\text{CCV}_{AB}, \text{CCV}_{BA}\} + N_1(\mu, \sigma^2) \\
\widehat{\text{CCV}}_{AC} &= \max\{\text{CCV}_{AC}, \text{CCV}_{CA}\} + N_2(\mu, \sigma^2) \\
\widehat{\text{CCV}}_{BC} &= \max\{\text{CCV}_{BC}, \text{CCV}_{CB}\} + N_3(\mu, \sigma^2)
\end{aligned} \tag{4.8}$$

where $N_x(\mu, \sigma^2)$, with $x = 1, 2, 3$ represents a similar internal noise as used in Equation 4.5 that correspond to three numbers drawn from a Gaussian distribution with $\mu = 0$ and $\sigma = 10.1$ MU. The pair having the maximum $\widehat{\text{CCV}}$ value was indicated by the artificial listener as the most similar pair. The pair having the minimum $\widehat{\text{CCV}}$ value was indicated as the least similar pair and, therefore, the remaining pair was indicated as having intermediate similarity. Since no external (ICRA) noise is used in the trials, the CCV values are deterministic but the responses of the artificial listener were not due to the internal (Gaussian) noise. To simulate the triadic comparisons of 20 participants, the 35 triads were evaluated 20 times by the artificial listener.

Table 4.2: Results of the simulations using a subset of 9 piano pairs and different t_{obs} durations. The minimum and maximum simulated thresholds are indicated together with their dynamic range (DR=$thres_{max}$−$thres_{min}$). The correlation values of the simulations with the corresponding experimental data (taken from Figure 3.4) are given. The SNR range of the experimental data is indicated in column Exp.

		"Observation (listening) period" t_{obs} [s]								
Parameter	Exp.	0.2	0.25	0.3	0.5	0.7	0.9	1.1	1.3	1.5
$thres_{max}$ [dB]	20.75	15.0	20.5	14.25	9.25	5.0	3.25	2.5	2.0	2.75
$thres_{min}$ [dB]	-1.75	-0.25	-1.0	1.5	-0.5	-1.25	-1.75	-2.75	-3.0	-2.5
DR [dB]	22.5	15.25	21.5	12.75	9.75	6.25	5.0	5.25	5.0	5.25
$r_p(7)$	–	0.66*	0.71*	0.65**	0.34	0.45	0.25	0.43	-0.21	-0.18
$r_s(7)$	–	0.60**	0.78*	0.47	0.11	0.49	0.21	0.49	0.09	0.03

(*) Significant correlation, $p < 0.05$. (**) Correlations that approach significance, $p < 0.10$.

4.5 Results

Each piano pair was tested in a separate instrument-in-noise experiment. The simulation results are compared with the corresponding experimental thresholds taken from Figure 3.4.

Experimental thresholds

The experimental thresholds of Figure 3.4 range between $thres_{exp,max}$= 20.75 dB (pair 23) and $thres_{exp,min}$= −1.75 dB (pair 26), having a dynamic range DR_{exp}=$thres_{exp,max}$−$thres_{exp,min}$= 22.5 dB. Since the pairs 23 and 26 are part of the subset of 9 piano pairs, this DR is also valid for the experimental thresholds using the subset of data.

4.5.1 Exploratory simulations

The results for the selection of 9 piano pairs are shown in Table 4.2. In the table, information about the minimum (lowest median) and maximum (highest median) estimated thresholds is shown. Their difference is indicated as the dynamic range (DR) in dB. The simulations that considered 1.5-s long piano sounds (whole duration of the sounds plus 0.2 s of silence) delivered thresholds between $thres_{sim,max}$= 2.75 dB and $thres_{sim,min}$= −2.5 dB with a DR of 5.25 dB. These thresholds are too low with respect to the experimental data. This means that the artificial listener has access to more information than the actual participants. As a way to remove available cues within the auditory model, the piano sounds were truncated to shorter durations. "Observation" durations t_{obs} of 0.20, 0.25, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, and 1.50 s were tested. The effective duration of the piano sounds is 0.10 s shorter, because of the initial silence in the waveforms. The t_{obs} durations between 0.9 and 1.5 s seem to have a constant DR of about 5 dB and for shorter durations,

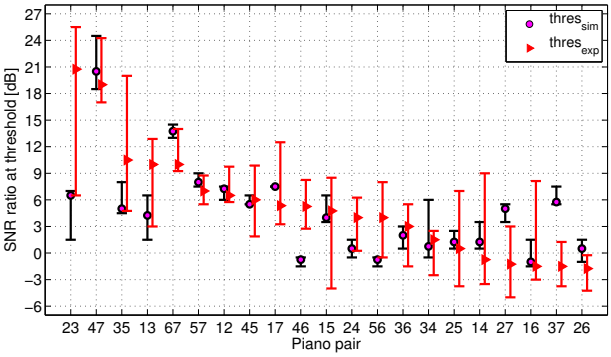


Figure 4.6: Discrimination thresholds using the whole dataset of piano sounds (21 piano pairs). The median simulated thresholds $\text{thres}_{\text{sim}}$ are indicated by the magenta circle markers. The red triangle markers correspond to the experimental thresholds $\text{thres}_{\text{exp}}$ (taken from Figure 3.4). The thresholds are shown together with their IQRs. The piano pairs along the abscissa are ordered from higher to lower SNR thresholds based on the experimental data.

the $\text{thres}_{\text{sim,max}}$ increases down to the duration of 0.25 s, reaching a maximum DR_{sim} of 20.5 dB. For the shortest tested duration of 0.20 s the DR decreases by 6 dB. The interpretation of these results is that at the duration of 0.25 s (that has the highest DR_{sim}) the piano sounds are judged by the auditory model as most distinct. Looking at the correlation values, the best fit between experimental and simulated data is found for the same observation duration of 0.25 s. For this duration, the thresholds have a Pearson correlation $r_p(7) = 0.71$, $p = 0.03$[8], and a Spearman (rank-order) correlation $r_s(7) = 0.78$, $p = 0.02$. This "observation" duration t_{obs} is further used to simulate the discrimination thresholds of the remaining 13 piano pairs.

4.5.2 Simulations using the whole dataset of piano sounds

The discrimination thresholds using the whole dataset of piano sounds (21 piano pairs) were simulated using the first 0.25 s of waveforms (i.e., initial 0.15 s of the piano sounds), based on the results of the exploratory simulations. The median thresholds $\text{thres}_{\text{sim}}$ are indicated by the magenta circle markers of Figure 4.6. The thresholds are shown together with their IQRs. The simulations at this duration ($t_{\text{obs}} = 0.25$ s) are not only highly correlated with the experimental data but they also reach a comparable DR_{sim}= 21.5 dB (same DR as in the exploratory analysis). The $\text{thres}_{\text{sim}}$ values range between $\text{thres}_{\text{sim,max}}$= 20.5 dB (pair 47) and $\text{thres}_{\text{sim,min}}$= −1 dB (pair 16). The Spearman (rank-order) correlation between the thresholds $\text{thres}_{\text{sim}}$ and $\text{thres}_{\text{exp}}$ is significant

[8]The value between brackets indicate the degrees of freedom, which is $N - 2$, with N being the number of data points being compared.

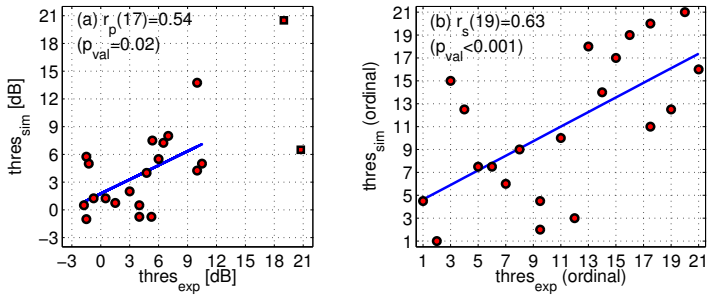


Figure 4.7: Regression analysis between the experimental $\text{thres}_{\text{exp}}$ and simulated $\text{thres}_{\text{sim}}$ as: (a) SNR thresholds, and (b) ordinal thresholds. The linear regression of panel (a) is related to the Pearson correlation r_p, while the regression of panel (b) to the Spearman correlation r_s. Two pairs of points were removed from the analysis to obtain an $r_p(17) = 0.54$, $p = 0.02$, due to the lack of $\text{thres}_{\text{exp}}$ values above 12 dB. A Spearman correlation of $r_s(19) = 0.63$, $p < 0.001$ was obtained.

with $r_s(19) = 0.63$, $p < 0.001$. Although a higher Pearson correlation $r_p(19) = 0.66$, $p < 0.001$ was found, this value has to be interpreted with caution due to the poor scattering of the $\text{thres}_{\text{exp}}$ values for SNRs above 12 dB[9]. When omitting the data of the two piano pairs in that range (pairs 23 and 47), a correlation $r_p(17) = 0.54$, $p = 0.02$ is found. The scatter plot of the data together with the corresponding regression analyses are shown in Figure 4.7. The poor scattering of the $\text{thres}_{\text{exp}}$ values can be seen in panel (a) of the figure, where there are only two thresholds in the SNR range between 12 and 24 dB (for pairs 23 and 47).

4.5.3 Comparison of the simulations with two perceptual MDS spaces

Euclidean distances from experimental triadic comparisons

The Euclidean distances $d_{ij\ \text{exp}}$ in the four-dimensional perceptual MDS space derived from the experimental results of the method of triadic comparisons have been taken from Figure 3.8. The first two dimensions of the space are replotted in panel (a) of Figure 4.8. The Euclidean distances range between $d_{ij\ \text{exp,min}} = 0.14$ (pair 47) and $d_{ij\ \text{exp,max}} = 0.91$ (pair 24). Half of the distances lie in the range $d_{ij,25-75} = 0.63 - 0.83$.

Euclidean distances from simulated triadic comparisons

The results of the triadic comparisons using the artificial listener, i.e., using the PEMO model, are shown in Table 4.3. In the table, the upper right triangle corresponds to the similarity matrix, which has been constructed in the same way as the matrix of Chapter 3. A four-dimensional

[9]The poor scattering of the data shown in Figure 4.7(a) is also related to the violation of the normality assumption of both, experimental and simulated thresholds.

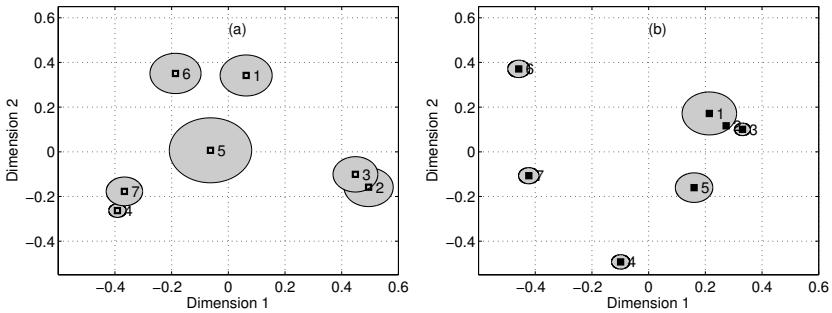


Figure 4.8: Perceptual spaces obtained with MDS. Only the first two (of four) dimensions are shown for the space constructed with a similarity matrix obtained experimentally (panel (a)) and with simulations (panel (b)). The grey bubbles give an indication of the "participant's" variability: the bigger the bubble the higher the variability across participants. Note that the axes of the MDS spaces are not to scale.

Table 4.3: Similarity matrix S_{ij} and Euclidean distances derived from the artificial listener using the test piano sounds. The similarity matrix is shown in the upper right triangle and the Euclidean distances between pianos in the resulting four-dimensional space are shown in the lower left triangle. To obtain these results, each triad was evaluated 20 times. This means that the maximum possible score is 200. The lowest score was obtained for pair 46 ($S_{ij} = 19$) and highest scores were obtained for pairs 13 and 23 (both with $S_{ij} = 167$). The corresponding distances were 0.95, 0.37, and 0.38 for pairs 46, 13, and 23, respectively.

	Piano						
Piano	P1	P2	P3	P4	P5	P6	P7
P1	-	145	167	75	130	77	83
P2	0.51	-	167	75	111	45	79
P3	0.37	0.38	-	86	162	59	61
P4	0.77	0.77	0.76	-	128	19	139
P5	0.57	0.65	0.38	0.57	-	49	102
P6	0.77	0.86	0.83	0.95	0.84	-	141
P7	0.76	0.77	0.80	0.51	0.70	0.51	-

perceptual space was obtained using the non-metric MDS algorithm available in MATLAB. The Euclidean distances between pairs in the fitted space are shown in the lower left triangle of Table 4.3. The obtained space has a goodness of fit that is near to excellent (stress $S_t = 3.6\%$) with respect to the similarity matrix, and its first two dimensions (poor stress $S_t = 25.8\%$) are shown in panel (b) of Figure 4.8. The Euclidean distances $d_{ij\ \mathrm{sim}}$ have Pearson and Spearman correlations of $r_p(19) = 0.51$ and $r_s(19) = 0.50$ (both with $p = 0.02$) with respect to the distances $d_{ij\ \mathrm{exp}}$. To further characterise the agreement between $d_{ij\ \mathrm{exp}}$ and $d_{ij\ \mathrm{sim}}$, a measure of stress (see Equation 3.3) can be used. Using $d_{ij\ \mathrm{exp}}$ as reference, the obtained stress is $S_{t\ \text{exp-sim}} = 25.2\%$. Additionally, the first dimension of both MDS spaces provide a similar rank order of the piano sounds with a Spearman correlation of $r_s(5) = 0.82$, $p = 0.03$.

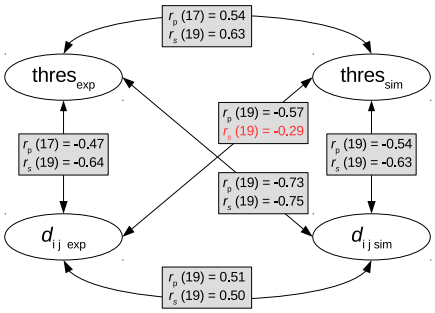


Figure 4.9: Summary of correlation values between instrument-in-noise thresholds and Euclidean distances. All possible combinations among thres$_{exp}$, thres$_{sim}$, $d_{ij\ exp}$, and $d_{ij\ sim}$ are indicated in this schema.

Euclidean distances and instrument-in-noise thresholds

In Chapter 3 a correlation value of $r_s(19) = -0.64$, $p = 0.001$ was reported for the distances $d_{ij\ exp}$ with respect to the instrument-in-noise thresholds thres$_{exp}$. The assessed correlation value between $d_{ij\ exp}$ and the simulated thresholds thres $_{sim}$ is a moderate value of $r_s(19) = -0.29$, $p = 0.20$. This indicates that the relationship between $d_{ij\ exp}$ and thres$_{sim}$ is less strong than with respect to thres$_{exp}$. When using the distances $d_{ij\ sim}$ as reference, the correlation values are $r_p(19) = -0.73$ and $r_s(19) = -0.75$ (both with $p < 0.001$) with respect to thres$_{exp}$ and $r_p(19) = -0.54$, $p = 0.01$, and $r_s(19) = -0.63$, $p = 0.002$ with respect to thres$_{sim}$. These values indicate a strong relationship between $d_{ij\ sim}$ and both, experimental and simulated thresholds.

All correlation values reported in this section are summarised in the schema of Figure 4.9.

4.6 Data analysis and discussion

The simulated thresholds thres$_{sim}$ of the instrument-in-noise test are significantly correlated with the experimental thresholds thres$_{exp}$ when only the initial part of the waveforms is used. Two aspects that affected the internal representation of the sounds leading to the obtained thres$_{sim}$ values are addressed in this section: (1) The weighting of information in each (audio and modulation) frequency channel, and; (2) the concept of "optimal detector" used in the central processor stage.

4.6.1 Weighting of information in the internal representations

The weighting of information for each audio and modulation frequency band within the PEMO model is inherently introduced when using the concept of memory template. This weighting depends on the "expected signal" to be discriminated and the processing introduced by each stage of the auditory model. Two aspects that may have affected the weighting of information in our approach are discussed: (1) the stronger limitation introduced in the adaptation loops, and (2) the processing of sounds that have shorter durations.

Our decision stage made use of two criteria, i.e., two expected signals that lead to two templates $T_{p,t}$ and $T_{p,r}$. Since the decision is based on CCV values, where the contribution of information to the difference representation ΔR_x is weighted using $T_{p,t}$ and $T_{p,r}$, the contribution of individual (audio and modulation) frequency bands to each CCV_x value can be assessed. Following a similar approach to that used to analyse the piano-alone representation R of piano P1 (shown in Figure 4.5), the weighting of information that is introduced within the auditory model can be obtained by assessing the percentual contribution of the template-weighted piano representations $\Delta R_x \cdot T_p$ (from Equation 4.3) using Equations 4.6 and 4.7. The contribution of each frequency band (I_m/I_{tot} for audio frequencies and I_k/I_{tot} for modulation frequencies) in the following conditions is considered: (1) when the adaptation loops are limited using a factor of 5 (as suggested in this thesis) and with a factor of 10 (as in the literature), and (2) considering the total duration of the piano-plus-noise sounds (1.5 s) and when only the first 0.25 s are evaluated. In this analysis, all 21 piano pairs were included. Since our interest is on the weighting of information at threshold, the difference representation $\Delta R = R - R_N$ is assessed at the noise level of the respective ICRA noise indicated by the simulated thresholds. The information-weighted values (I_m/I_{tot} and I_k/I_{tot}) for the comparison between limiter factors are shown in Figure 4.10. The values I_m and I_k were obtained as the median of 42 values (21 pairs with one value using $T_{p,t}$ and one value using $T_{p,r}$). The error bars indicate their IQRs. The weighting I_m/I_{tot} shown in panel (a) shows that by using a stronger limiter factor, the information of higher audio frequency bands receive a higher relative weighting. For a limiter factor of 10, the weighting seems to be very similar to the distribution of information for the piano-alone representation shown in panel (a) of Figure 4.5. The information contribution of

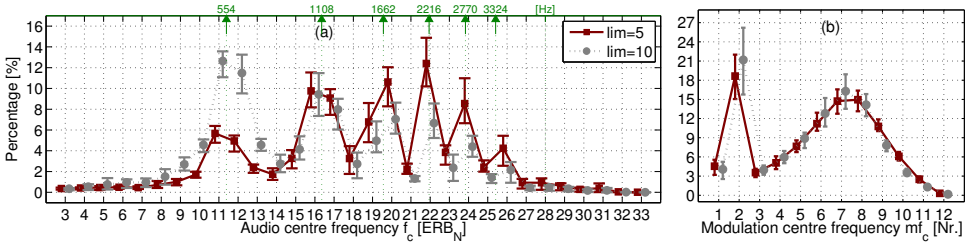


Figure 4.10: Weighting of information in difference (internal) representations ($\Delta R_x \cdot T_p$) for limiter factors of 5 (maroon square markers) and 10 (grey circle markers). The weighting I_m/I_{tot} of each audio frequency channel is shown in panel (a). The weighting I_k/I_{tot} of each modulation frequency channel is shown in panel (b). The values per band are expressed as percentage with respect to the total information I_{tot} of each representation. The points along the ERB scale that correspond to F0= 554 Hz and its five first harmonics are indicated by the green numbers along the top axis.

each modulation filter is shown in panel (b) of Figure 4.10. The average information contribution in the second modulation filter (mf_c = 5 Hz) is 18.6% for the representations with limit= 5, which is 2.6% below the weighting for the same filter when the limit of 10 is used (weighting of 21.2%). The first modulation filter has a low weighting despite its higher value of information content in the piano-alone representation. This is expected, though, because this modulation filter tracks slow envelope changes that do not differ considerably from piano to piano, especially in the first 0.25 s of their representation. The modulation filters Nr. 6-9 show a slight increase in their weighting for a limit of 5 (compared to the limit of 10), while the rest of the bands have a similar weighting with both limiter factors. The information-weighted values for the comparison between signal durations are shown in Figure 4.11. The band weightings using t_{obs} durations of 0.25 and 1.5 s are very similar (mean difference $\Delta I_m/I_{\text{tot}}$ of 0.00%, IQR of 0.33%) and, therefore, they seem to be unaffected by the duration t_{obs} of the piano sounds. To explain the influence of t_{obs} on the simulated thresholds, the performance of the artificial listener is further analysed in the next section.

4.6.2 Reducing the performance of the optimal detector

The central processor of the PEMO model is inspired by the concept of "optimal detector". In signal detection theory, the term "optimal" refers to the fact that the detector has the best possible performance given specific stimulus properties. In other words, if a cue is available in the stimulus, then the detector uses it (Green & Swets, 1966, their Chapter 6). For this reason, detectors that are optimal can be used as baselines for human detection. The results of our exploratory simulations showed

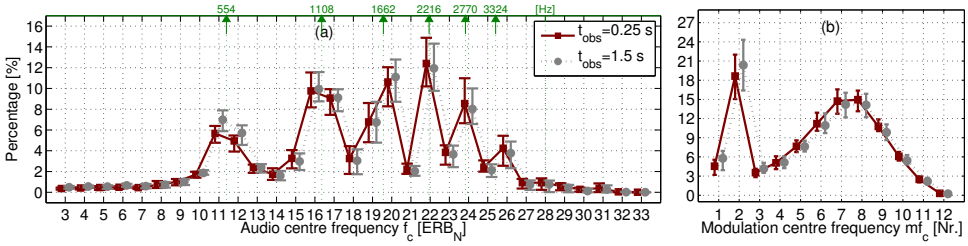


Figure 4.11: Weighting of information in difference (internal) representations ($\Delta R_x \cdot T_p$) for whole-duration sounds (grey circle markers) and considering a shorter duration t_{obs} of 0.25 s (maroon square markers). The weighting I_m/I_{tot} of each audio frequency channel is shown in panel (a). The weighting I_k/I_{tot} of each modulation frequency channel is shown in panel (b). The points along the ERB scale that correspond to F0= 554 Hz and its five first harmonics are indicated by the green labels on the top axis.

that the participant's performance in the instrument-in-noise experiment is below the ideal performance, where the simulated thresholds for whole-duration sounds covered a range of only 5 dB (see the last two columns of Table 4.2).

One way to bring the simulated thresholds to a range closer to that of the experimental data is the removal of "evidence" from the stimuli. Since the "evidence" is assumed to be accumulated during the observation period, shortening the duration of the sounds should result in a reduction of evidence and an increase in simulated thresholds.

Shortening the duration of the sounds

We systematically varied the observation period t_{obs} of the artificial listener by only considering the initial part of the piano (and ICRA noises) waveforms, which was indicated by the t_{obs} duration. The simulated thresholds for shorter t_{obs} durations resulted in thresholds with a higher dynamic range ($\text{thres}_{\text{max}} - \text{tres}_{\text{min}}$), increasing from 5.25 dB for $t_{\text{obs}} = 1.5$ s to 21.5 dB for $t_{\text{obs}} = 0.25$ s.

To evaluate the influence of different t_{obs} periods on the decision of the artificial listener, an analysis based on CCV values is presented. For this analysis, the optimal and the longest t_{obs} periods of 0.25 and 1.5 s, respectively, are used. The CCV values for the subset of 9 piano pairs are shown at a noise level given by their corresponding $\text{thres}_{\text{sim}}$ value. In general, at these noise levels only one of the two decision criteria fails, either $\max\left\{\widehat{\text{CCV}}_{x,t}\right\}$ or $\min\left\{\widehat{\text{CCV}}_{x,r}\right\}$ (see Equation 4.3). The criterion that fails first is labelled as "leading criterion" and is used for further analysis.

The CCV values for the selected piano pairs obtained at the corresponding discrimination threshold ($\text{thres}_{\text{sim}}$ using $t_{\text{obs}} = 0.25$ s) are shown

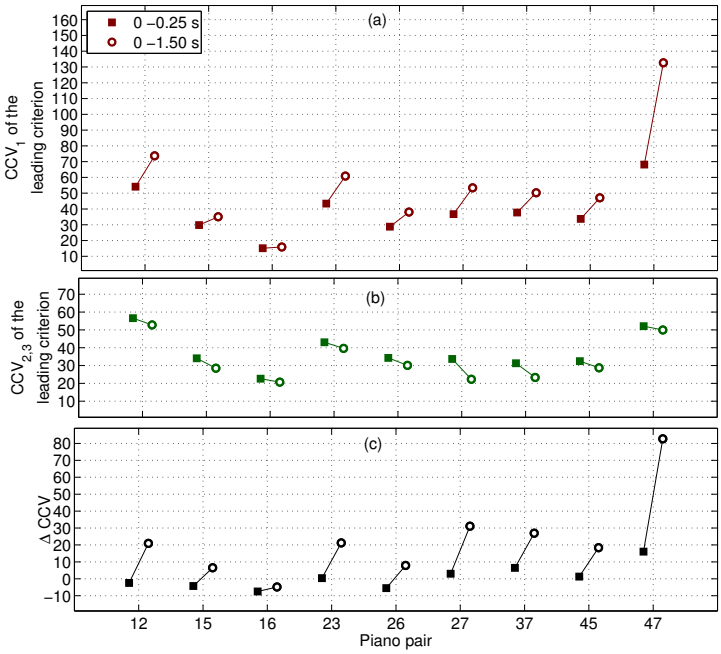


Figure 4.12: CCV values for each piano pair (SNR at threshold) considering the first 0.25 s (filled square markers) and the whole duration (open circle markers) of the internal representations. In panel (a) the CCV values for the target interval ($x = 1$) of the leading criterion (CCV_t or CCV_r) are shown. In panel (b) the CCV values for the reference interval (valid for $x = 2$ and 3) of the same criterion (CCV_t or CCV_r, respectively) are shown. In panel (c) the difference between CCV values are shown.

in Figure 4.12. This means that, after adding internal noise $N(0, \sigma^2)$, the CCV values obtained from representations with $t_{\text{obs}} = 0.25$ s (filled square markers) would lead the artificial listener to obtain discrimination scores of approximately 70.7%[10]. The CCV values obtained from representations with $t_{\text{obs}} = 1.50$ s are indicated by open circle markers. The CCV values of the leading criterion for target and reference intervals are shown in panels (a) and (b) of the figure, respectively. The difference between CCV values is shown in panel (c) and they range between −7.5 (pair 16) and 16.0 (pair 47) for representations with $t_{\text{obs}} = 0.25$ s and between −4.8 (pair 16) and 82.7 (pair 47) for representations with $t_{\text{obs}} = 1.50$ s. These ΔCCV values indicate that the discriminability of the pianos either remains approximately constant (pair 16) or increases (pairs 12, 15, 23, 26, 27, 37, 45, and 47) with t_{obs} and that the use of shorter internal representations compresses the $\Delta CCV_{0.25\text{ s}}$

[10]For this CCV analysis no level roving was applied. This means that the discriminability at $\text{thres}_{\text{sim}}$ is in practice higher. This is due to, on average, the lower thresholds (i.e., better discriminability) when removing the level roving, as can be seen in no-rove thresholds of Figure 4.13.

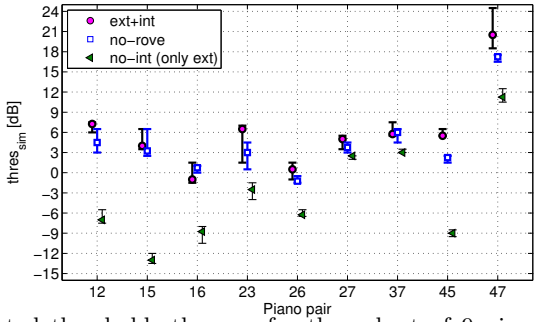


Figure 4.13: Simulated thresholds $\text{thres}_{\text{sim}}$ for the subset of 9 piano pairs in the following conditions: (1) considering internal and external sources of variability (magenta circle markers, as in Figure 4.6); (2) with internal variability but without level roving (square blue markers); (3) without internal variability, i.e., considering only sources of external variability.

without changing significantly the relative discriminability between pianos, having a rank-order correlation of $r_s(19) = 0.93$, $p < 0.001$ with respect to the $\Delta\text{CCV}_{1.50\text{ s}}$ values. The differences $\Delta\text{CCV}_{0.25\text{ s}}$ are, however, susceptible to the variance introduced by the internal noise. Since each CCV value is varied by a number drawn from a normal distribution having the same standard deviation $\sigma = 10.1$ MU, the difference ΔCCV values are also normally distributed with a standard deviation of $\sqrt{\sigma^2 + \sigma^2} = 14.4$ MU. Eight of the 9 difference $\Delta\text{CCV}_{0.25\text{ s}}$ values in panel (c) of Figure 4.12 (20 of 21 if the whole dataset is considered) lie in the variability range of the internal noise ($\pm$14.4 MU). This means that the internal noise plays a prominent role in the discrimination performance of the artificial listener. For representations with $t_{\text{obs}} = 1.5$ s a much larger variance of the internal noise would be needed for reaching simulated thresholds in a similar SNR range. Although it is possible to introduce a higher variability to the internal representations, this would strongly limit the performance of the PEMO model, reducing its sensitivity when predicting auditory tasks like those shown in Appendix D.

4.6.3 Removing the sources of variability

In order to quantify the influence of the sources of variability on the obtained thresholds $\text{thres}_{\text{sim}}$, simulations for the subset of 9 piano pairs (using $t_{\text{obs}} = 0.25$ s) were run in the following conditions: (1) No level roving (no–rove condition), i.e., using only the internal noise variability and the use of running noises, and (2) No internal noise (no–int condition), i.e., using only sources of external variability (level rove and running noise). The resulting median thresholds (of 6 estimates) with

their IQRs are indicated by the blue squares and the green triangles in Figure 4.13, respectively. The simulated thresholds using both sources of variability (as shown in Figure 4.6) are indicated by the magenta circle markers (ext+int condition) and are used as baseline for this analysis. The simulated thresholds in the no–rove condition follow the trend of the ext+int-thresholds (correlation of $r_s(7) = 0.77$, $p = 0.02$) and differ by 3.5 dB (pair 23) or less. This is not the case for the thresholds in the no–int condition, that are much lower than the ext+int-thresholds and are not significantly correlated ($r_s(7) = 0.53$, $p = 0.15$). This means that the limit in performance introduced by the sources of external variability of the instrument-in-noise task are not sufficient to explain the performance of the artificial listener. This analysis provides evidence of the dominant role played by the internal noise in the decision of the artificial listener for 0.25-s long representations.

4.6.4 Comparison between simulated thresholds and simulated perceptual distances

Although the Euclidean distances $d_{ij\ \mathrm{exp}}$ and instrument-in-noise thresholds $\mathrm{thres}_{\mathrm{exp}}$ obtained in Chapter 3 have a high correlation ($r_p = -0.47$, $r_s = -0.64$), and the correlation between simulated thresholds $\mathrm{thres}_{\mathrm{sim}}$ obtained in this chapter have a high correlation with $\mathrm{thres}_{\mathrm{exp}}$ ($r_p = 0.54$, $r_s = 0.63$), the distances $d_{ij\ \mathrm{exp}}$ have a moderate to low correlation with $\mathrm{thres}_{\mathrm{sim}}$ ($r_p = -0.57$, $r_s = -0.29$). For understanding why the discrimination thresholds (in noise) of the PEMO model are not better correlated with the experimental results of the triadic comparisons $d_{ij\ \mathrm{exp}}$, we integrated the triadic comparison task into the framework of the auditory model. The simulated distances $d_{ij\ \mathrm{sim}}$ had a similar strength of association with both, the distances $d_{ij\ \mathrm{exp}}$ ($r_p = 0.51$, $r_s = 0.50$) and the thresholds $\mathrm{thres}_{\mathrm{sim}}$ ($r_p = -0.54$, $r_s = -0.63$). An interpretation of these results can be that the artificial listener does not fully perceive the similarity of piano sounds in silence in the way participants do. This is evidenced by the poor stress between distances $S_{t\ \mathrm{exp\text{-}sim}} = 25.2\%$, while the stress values between distances and their corresponding similarity matrices are between good and excellent. The non-explained variance of the dimensions in the simulated MDS space (with the experimental space) seems to be responsible for the better correlation between $d_{ij\ \mathrm{sim}}$ and $\mathrm{thres}_{\mathrm{sim}}$.

4.7 Conclusions

In this chapter an auditory model was used to simulate the discrimination thresholds between recorded sounds of one note ($C\#_5$) played on 7 different pianos. In order to compare two internal (piano) representations, two memory templates were required to allow the artificial listener (PEMO model) to distinguish one piano from another. The need of the model to access the representation of the sounds being compared can be interpreted as an approach that resembles a recognition rather than a discrimination task. The obtained $\text{thres}_{\text{sim}}$ values from the model were significantly correlated with the $\text{thres}_{\text{exp}}$ values when only the initial part of the waveforms was used. An optimal "observation" duration t_{obs} of 0.25 s was found. We hypothesise however that other t_{obs} durations will be obtained if other (piano) notes are tested. The relevant aspect is that a reduction in the amount of information available to the artificial listener brought the simulated and experimental data to a closer range. In this context, the success of the simulations might be interpreted in the following way: (1) Using longer t_{obs} durations, the artificial listener has access to more cues than the actual participants. This may be related to the fact that the central processor integrates the incoming information "optimally"; (2) the shorter the t_{obs} duration the less information can be integrated by the artificial listener. The performance of the artificial listener is limited by the internal noise of the auditory model and by other sources of (external) variability that are related to the instrument-in-noise task. These sources of external variability are the randomisation of the presentation level of each interval (level roving) and the use of running ICRA noises. The most dominant of the sources of variability is the internal noise in the auditory model, especially for intervals with a t_{obs} of 0.25 s and SNR levels around or above the simulated thresholds. The results presented in this chapter support the idea that the unified framework offered by the PEMO model can be used to evaluate perceptual tasks using complex sounds. This can be seen as an extension of the use of this type of models and their success relies on the adjustment of the central processor stage included within the model, in combination with an appropriate representation of sources of internal noise.

5 Measuring and simulating the similarity of instrument sounds in a reverberant environment[1]

5.1 Introduction

The perceptual similarity task studied in the previous chapters is applied here to the same set of piano sounds but after being convolved with the impulse response of a reverberant room. Although the judgements for the new reverberant sounds are expected to be somewhat correlated with the results reported in Chapter 3, relative similarity changes among pianos due to reverberation are expected to be tracked by the instrument-in-noise method and by simulations using the auditory (PEMO) model described in Chapter 4. One of the objectives of the study case presented in this chapter is, therefore, to extend the use of the experimental and computational frameworks presented in Chapters 3 and 4. The experimental data of the instrument-in-noise method using reverberant sounds are compared with the method of triadic comparisons and with simulations of the discrimination thresholds using the PEMO model. As a difference to the procedures of the previous chapters, a new version (version B) of the ICRA noise algorithm has been adopted. Version B of the algorithm corrects the spectral tilt towards high frequencies that the algorithm version A had (see Section 3.5). For this reason, a second objective of this chapter is to quantify the effect of using ICRA noises with different spectral properties. The evaluation is done by comparing simulated discrimination thresholds using the two ICRA noise versions.

5.2 Description of the method

The experimental methods and the computational framework used in the study case presented in this chapter are very similar to those used in Chapters 3 and 4, respectively. The set of stimuli comprises the same

[1] This chapter is partly based on: A. Osses, and A. Kohlrausch (2018, submitted). Auditory modelling of the perceptual similarity between piano sounds. *Acta Acust. united Ac.*

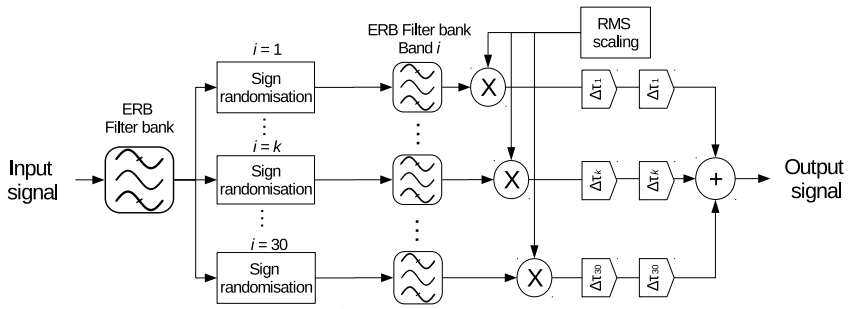


Figure 5.1: The principle of the ICRA noise generation, version B. For details in the procedure, refer to steps 1 to 4 in the text.

19th-century Viennese pianos, but the sounds were auralised to account for the acoustics of a reverberant room. The ICRA noises used to mask the auralised piano sounds have been calculated with a modified algorithm, whose resulting waveforms are more similar to the outputs of a 30-channel noise vocoder. Since the description of the experimental sessions and simulations is, in general, shorter than the descriptions of the previous chapters, the reader is referred to Chapters 3 and 4 for specific details about the procedures.

5.2.1 Modified ICRA noise, version B

The procedure used to generate ICRA noises has some modifications with respect to the algorithm version A (Section 3.2.1). The modified ICRA algorithm, that has been named "version B", is shown in the block diagram of Figure 5.1 and can be described as follows:

1. **Band-split filter**: an input signal (musical instrument sound) is fed into a Gammatone filter bank. The filter bank consists of 30 bands with centre frequencies between 101 Hz (3.4 ERB_N[2]) and 7324 Hz (33.4 ERB_N), spaced at 1 ERB. This number of bands was obtained by using F0= 554 Hz (11.4 ERB_N) as base frequency. The all-pole Gammatone filter bank with complex outputs (only the real part is further processed) available in the AMT toolbox for MATLAB was used for this purpose. The filter design and processing introduced in this stage is equivalent to the "frequency analysis" stage described by Hohmann (2002).

2. **Sign randomisation**: the sign of each sample of the 30 filtered signals is either reversed or kept unaltered with a probability of 50% (multiplication by 1 or −1) (Schroeder, 1968). As a consequence of this

[2]The ERB rate scale corresponds to one of the frequency scales that is inspired by the frequency representation in the auditory system. A brief overview of this scale is given in Appendix A.

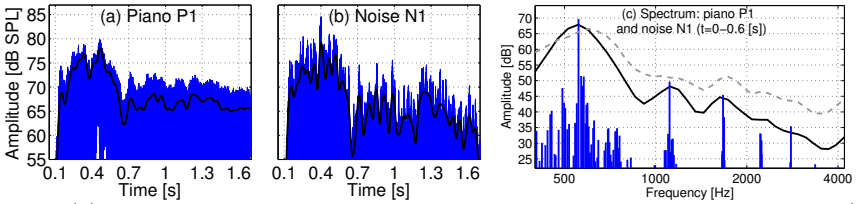


Figure 5.2: (a) Waveform of a reverberant sound of piano P1 converted to SPL, and (b) one realisation of its resulting ICRA noise at an SNR= 0 dB. The thick black lines correspond to the envelope of the waveforms (LPF, $f_{\text{cut-off}} = 20$ Hz). (c) Spectra of the piano sound (blue) and the ICRA noise (black thick line) averaged over the first 0.6 s of both waveforms. The grey dashed line represents the spectrum of the ICRA noise, using the old version A.

process, the resulting waveforms have a flat spectrum while keeping the same temporal envelope characteristics and the same band level.

3. **Re-filtering per band-split filter**: the resulting signal from band i is fed into the ith band of the Gammatone filter bank. The index i represents each of the 30 bands. As a consequence of this process, the band levels are decreased in proportion to the number of rejected frequency components. To compensate this effect, a gain is applied to set the band levels back to the values as before this stage.

4. **Add signals together**: the 30 filtered signals are added together. A frequency dependent delay line is used before adding the filtered signals together. This is because the Gammatone filter bank is implemented as a set of IIR filters and, therefore, the filter bank has frequency-dependent group delays. The delays being compensated range from 5.6 ms (bands centred at $f_c = 554$ Hz or below) down to 0.57 ms (band centred at $f_c = 7324$ Hz). Those delays correspond to the time stamp at which each BPF ($f_c \geq 554$ Hz) has a maximum in its envelope, when an impulse is used as input. For the filters with $f_c < 554$ Hz only a partial compensation (of 5.6 ms) is applied. The processing introduced in this stage is similar to the "frequency synthesis" stage described by Hohmann (2002) but omitting the fine-structure alignment. This compensation is applied twice (two-tap delay line) because the signals are filtered (stage 1) and then re-filtered (stage 3).

Difference between versions A and B

The level scaling introduced in the current ICRA algorithm (version B) ensures a resulting noise that has the same overall level per critical band as the input signal. In version A, the level is only adjusted after the noise has been summed up into one broadband signal. This means that both

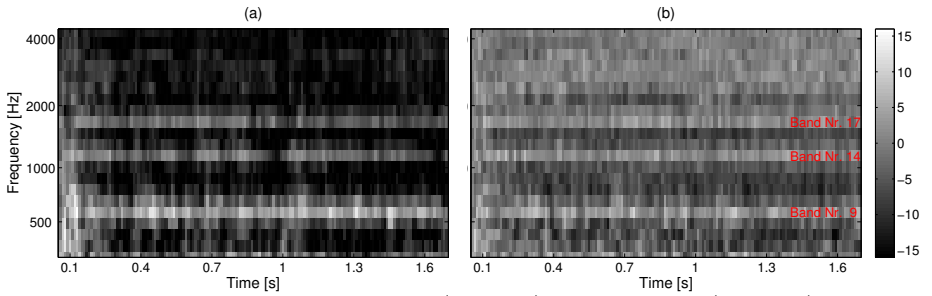


Figure 5.3: SNR map as a function of time (abscissa) and frequency (ordinate) for piano P1 with respect to noise N1 at an SNR= 0 dB. Noise N1 was obtained using version A (panel (a)) and B (panel (b)) of the ICRA noise algorithm. The overall SNR between P1 and N1 averaged across time and frequency for both noises is: −12.1 dB for version A and −0.5 dB for version B.

versions deliver noises with the same overall level as the input signal, but in version A the band levels show a spectral tilt towards higher frequencies. This is a consequence of the re-filtering stage, where the band levels after the signal randomisation stage (which are not changed with respect to the levels before this stage) are decreased in inverse proportion to the bandwidth of the critical band. As a consequence of this, the band levels are less attenuated for higher frequency bands in version A[3] (the relative level of the last auditory filter is emphasised by 10 dB with respect to the band level in the auditory filter centred at F0= 554 Hz). In panel (c) of Figure 5.2 the band levels of the ICRA noise versions B (black solid line) and A (grey dashed line) are shown. To further characterise the differences between versions A and B of the ICRA algorithm, the SNR maps of Figure 5.3 have been drawn, were darker and brighter regions indicate lower and higher SNRs, respectively. Those maps show the SNR as a function of time and frequency between piano P1 and two ICRA noise realisations obtained from versions A (panel (a), as in previous chapters) and B (panel (b), as used in this chapter), respectively. For both ICRA noises, the bands containing the F0 and the first two harmonics have positive SNRs (bands 9, 14, 17). The levels

[3] At this point of the ICRA algorithm, the re-filtering stage keeps the spectrum level of the white-noise like waveforms. If each auditory filter contains a signal with a band level BL_i with a wideband spectrum $\mathrm{BW}_{\text{full-range}} = f_s/2$ Hz, then the spectrum level of the band is $\mathrm{SL}_i = \mathrm{BL}_i - 10 \cdot \log_{10}(f_s/2)$. After the re-filtering the signals are limited (as before Stage 2) to the bandwidth BW_i of the corresponding Gammatone filter, then $\mathrm{BL}_{i\ \text{new}} = \mathrm{SL}_i + 10 \cdot \log_{10}(\mathrm{BW}_i)$, with $\mathrm{BL}_{i\ \text{new}}$ being a level that is always lower than BL_i. By construction, the attenuation introduced by the re-filtering stage is given by $\mathrm{Att}_i = 10 \cdot \log_{10}(\mathrm{BW}_{\text{full-range}}/\mathrm{BW}_i)$. For an $f_s = 44100$ Hz, the values Att_i for the band centred around F0= 554 Hz ($\mathrm{BW}_i \approx 70$ Hz) and in the highest auditory band $f_c = 7324$ Hz ($\mathrm{BW}_i \approx 700$ Hz) are 25.0 and 15.0 dB, meaning that the BL_i of the highest band has higher inherent weighting (by 10 dB) over the band level in the band centred at F0. In version B Att_i values are compensated but in version A they are not.

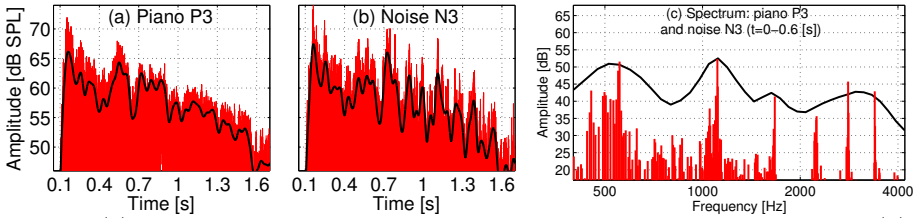


Figure 5.4: (a) Waveform of a reverberant sound of piano P3 converted to SPL, and (b) one realisation of its resulting ICRA noise at an SNR= 0 dB. The thick black lines correspond to the envelope of the waveforms. (c) Spectra of the piano sound (red) and the ICRA noise (black thick line) averaged over the first 0.6 s of both waveforms.

of the auditory bands in between have, on average, negative SNRs, with lower values for the map that uses the ICRA noise version A (panel (a)) with SNRs below −10 dB. Another aspect to highlight in the SNR map that uses the ICRA noise version A is that there are high SNRs (brighter region) between about 40 ms before P1 starts and up to about 30 ms after its note onset (between $t = 60$ and 130 ms) for band 11 or below ($f_c \leq 743$ Hz). The noise has levels below the signal in that range (i.e., SNR$>$ 0 dB) due to the frequency-dependent delay (group delay) which is longer for lower frequency bands. The group delay is introduced by the ICRA noise algorithm. In the current "version B" the group delay compensation seems to have solved that problem.

5.2.2 Comparing two sounds

Two sounds are compared by measuring the participant's discrimination performance using background ICRA noises in exactly the same way as used in the previous chapters (see Section 3.2.2). To explain the comparison procedure, two recordings of the note C#$_5$ from pianos P1 and P3 are used (see Table 5.1). Since the piano sounds used in this chapter include the effect of reverberation, the spectro-temporal properties of the piano sounds vary from those of the waveforms used in previous chapters. The generated ICRA noises follow these variations. The chosen reverberant sounds together with one realisation of their ICRA noise (at an SNR of 0 dB) are shown in Figures 5.2 and 5.4.

Practical considerations

During the experimental procedure, a 3-AFC discrimination task is used to compare two sounds. The sounds being compared are set to have the same duration of 2.0 s. This increased duration (1.3 s was used in the previous dataset) is assumed to be long enough to convey all the relevant cues that the reverberation may introduce onto the piano

sounds[4]. The piano onset is set to occur at approximately the same time stamp ($t = 0.1$ s).

In order to simultaneously account for the spectro-temporal properties of two piano sounds (e.g., P1 and P3), a "paired" ICRA noise is generated by averaging the waveforms of two individual ICRA noises (N1 and N3). It is assumed that the paired ICRA noise (N13) is efficient to gradually mask the properties of the test sounds (P1 and P3) when being compared to each other. In the course of an experiment, twelve realisations of a paired ICRA noise are used, where three realisations are randomly chosen for each trial. This corresponds to an approximation to a running-noise condition. The relative level of the paired noises is adapted in the course of the experiment by increasing the level of the noise (decrease of the SNR, more difficult discrimination) or decreasing the level of the noise (increase of the SNR, easier discrimination), depending on the participant's performance.

5.2.3 Instrument-in-noise test

The instrument sounds are compared pairwise. A given pair of sounds is presented in 3-AFC trials, where the discriminability threshold is estimated by adjusting the noise level of the corresponding paired ICRA noise, version B. The participant has to indicate which of the intervals contains the target sound. The adjustable parameter (noise level) is varied following a two-down one-up rule. The experiment continues until 12 reversals are reached. The starting level of the paired ICRA noise is set to an SNR of 16 dB. The step size at which the noise is adjusted is set to 4 dB and is halved after two reversals until a step size of 1 dB is reached. The median of the last 8 reversals is used to estimate the discrimination threshold of each pair of sounds.

The reverberant piano sounds used in this chapter differ considerably in their loudness. In order to avoid the use of loudness cues during the experimental sessions, the piano sounds were first loudness balanced ($S_{\max}$ set to approximately 18 sone, as shown in Table 5.1) and their presentation level within each interval (piano + noise) was randomly varied (roved) by levels in the range ± 4 dB, drawn from a uniform distribution. The intervals had a duration of 2.0 s with an interstimulus interval of 0.2 s. A similar balanced subset of data, as used in Chapter 3, was considered for each participant with the goal of reducing the duration of the

[4]This may be a strong assumption given that the reverberation time (RT) of the acoustic space studied in this chapter is longer than 2 s (as shown later in Table 5.2). The initial (practical) motivation of this "short" stimulus duration is to limit the duration of the experimental sessions.

experimental sessions. In this way one evaluation of the whole dataset (21 pairs) was obtained every two participants.

5.2.4 Reference procedure: Triadic comparisons

The method of triadic comparisons was used to obtain similarity judgements between stimuli. Within a trial, three sounds (A, B, C) are presented and the participant is asked to indicate which of the three possible pairs (AB, AC, BC) contains the most similar sounds and which one contains the least similar sounds. These judgements are counted and summarised in a similarity matrix. The results of the similarity matrix are further processed using the MDS algorithm, where the stimuli are mapped onto a q-dimensional space. The Euclidean distances between stimuli within the resulting space correspond to a unidimensional measure of similarity that is used as the reference to be compared with the discrimination thresholds of the instrument-in-noise test.

5.2.5 Instrument-in-noise test: Simulations

The simulations consider the implementation of the instrument-in-noise test in the same way as in the actual experimental sessions, but using only the left-ear channel of the sounds. This limitation is imposed by the use of a monaural auditory model and it assumes that the right-ear signal would lead to a similar performance within the auditory model.

The simulation is then implemented as an adaptive 3-AFC experiment, where discriminability thresholds expressed as SNRs in dB are estimated. Each staircase simulation is stopped after 8 reversals. This means that the threshold estimates are based on the median value of 4 reversals at which the noise level is adapted in steps of 1 dB. This decision was made in order to reduce the time required to run the simulations. Exploratory simulations were first run using a subset of 9 piano pairs to test different t_{obs} durations in a similar way as done in Chapter 4. The t_{obs} duration that lead to the best fit between the simulated thresholds and the corresponding experimental thresholds was then used to simulate the thresholds $\text{thres}_{\text{sim}}$ using the whole dataset of pianos (21 piano pairs). These simulations were run using ICRA noises version B, as used in the experimental sessions. A final set of simulations was run to estimate simulated thresholds $\text{thres}_{\text{sim,A}}$ using ICRA noises version A. The aim of this last set of simulations was to quantify how much the $\text{thres}_{\text{sim}}$ values deviate from the thresholds estimated using ICRA noises version A (as used in Chapters 3 and 4).

Table 5.1: List of pianos and level information of their auralised sounds as used in the listening experiments. The loudness of the sounds when presented 4 dB softer and 4 dB harder are shown in parentheses.

		Level [dB SPL]	Loudness [sone]
ID / Year	Manufacturer	L_{max} / L_{eq}	S_{max} / S_{avg}
P1 / 1805	Gert Hecher	80.0 / 67.3	17.3 (13.6-22.0) / 8.8 (6.7-11.5)
P2 / 1819	Nannette Streicher	74.4 / 59.2	16.9 (13.3-21.4) / 6.7 (5.0- 8.8)
P3 / 1828	Conrad Graf	73.1 / 55.8	17.1 (13.4-21.6) / 6.9 (5.1- 9.1)
P4 / 1836	Johann B. Streicher	78.6 / 64.7	17.1 (13.4-21.8) / 8.6 (6.5-11.2)
P5 / 1851	Johann B. Streicher (English)	77.5 / 62.9	17.0 (13.4-21.5) / 7.1 (5.5- 9.2)
P6 / 1851	Johann B. Streicher (Viennese)	81.0 / 68.1	18.0 (14.1-22.8) / 8.6 (6.5-11.2)
P7 / 1873	Johann B. Streicher & Sohn	80.9 / 69.8	17.4 (13.6-22.1) / 10.1 (7.7-13.1)

Table 5.2: Reverberation time in octave bands derived from the selected BRIR (AIR database, Aula Carolina, distance source-receiver of 4 m, azimuth of 90^o).

	Frequency [Hz]						
	125	250	500	1000	2000	4000	500/1000
T_{20} [s]	9.0	6.4	3.9	3.1	2.6	1.8	3.5
EDT [s]	6.5	5.8	3.4	2.6	1.8	1.0	3.0

5.2.6 Stimuli

The same set of recordings obtained from 19th-century Viennese pianos of the previous chapters are used in this chapter but the sounds were digitally auralised to account for the acoustics of a room. The sounds are, therefore, recordings of one note (C#$_5$, F0= 554 Hz) from seven pianos. The BRIR used for the auralisations corresponds to an existing measurement of Aula Carolina, which is a former church located in Aachen (Germany) that has a ground area of 570 m^2 and a high ceiling. The selected BRIR corresponds to an existing measurement done at a distance of 4 m and azimuth of 90^o with respect to the sound source and it was retrieved from the AIR database[5]. The estimated early decay time (EDT) is 3.0 s at mid frequencies (see Table 5.2). After auralising the piano sounds using digital convolution, the duration of each sound was set to 2.0 s, with the note onset occurring at a time stamp of 0.1 s. Some information about the resulting piano sounds is shown in Table 5.1. The sounds were ramped down using a 300-ms linear ramp. The loudness of the sounds was adjusted to have a maximum value of approximately 18 sone. For that purpose the short-term loudness from the TVL model (Glasberg & Moore, 2002) was used. After the adjustment, the individual piano sounds had a maximum level ranging from 73.1 to 81.0 dB SPL.

[5]AIR database (retrieved on 17/03/2017): http://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen-impulse-response-database/. Last accessed on 18/07/2018.

5.2.7 Apparatus

The experiments were conducted in a doubled-walled sound-proof booth. The stimuli were presented via Sennheiser HD 265 Linear circumaural headphones in a binaural reproduction. The participant's responses were collected on a computer using the software APEX (Francart et al., 2008) and the APE Toolbox for MATLAB (De Man & Reiss, 2014) for the instrument-in-noise and the triadic comparisons, respectively. The simulations were run using the AFC toolbox (Ewert, 2013) where it is possible to enable the use of an "artificial listener". The artificial listener uses the PEMO model with the same central processor as used in Chapter 4.

5.2.8 Participants

Twenty participants (3 females and 17 males) were recruited from the JF Schouten subject database of the TU/e university. At the time of testing, the participants were between 19 and 66 years old[6] (average of 26) and they all had self-reported normal hearing. They provided their informed consent before starting the experimental session and were paid for their contribution.

The sample size of 20 participants was assessed a priori aiming at testing the hypothesis that the data from the instrument-in-noise are highly correlated (effect size or Pearson correlation r_p of at least 0.6) with the data from the triadic comparisons, with a power of 90%. This analysis was done in the software G*Power (Faul et al., 2007, 2009), requiring 17 participants to reach the desired effect size. By increasing the number of participants to 20 the observable effect size is reduced to 0.57.

5.2.9 Data collection: Experimental sessions

The experimental sessions were organised in a similar way as in the experiment reported in Chapter 3. There were two one-hour sessions per participant, including breaks. For the instrument-in-noise test, every participant tested 10 or 11 piano pairs meaning that from every two participants one threshold estimate of the whole dataset (21 piano pairs) was obtained. The participants started the first session with the evaluation of 17 randomly chosen triads, followed by 5 threshold estimations (staircase procedure). During the second session the participants evaluated

[6] With the exception of one participant aged 66 years, all participants were between 19 and 26 years old at the time of testing. Their hearing thresholds were not measured but we assumed a normal hearing condition. The participant aged 66 year, however, may have had some hearing loss but since all his staircases met at least one of the data exclusion criteria, his data were not further used.

the remaining 18 triads, followed by other 5 or 6 threshold estimations, completing the total of 10 or 11 estimations.

5.2.10 Data collection: Simulations

Exploratory simulations: Subset of piano sounds

As done in Chapter 4, first a subset of 9 out of the 21 available piano pairs was used for the simulations. This subset was used to find the duration of the "observation" period t_{obs} of the artificial listener that provides the highest correlation value between the corresponding simulated and experimental thresholds. The durations t_{obs} of 0.16, 0.20, 0.25, 0.3, 0.5, 0.8, 1.0, 1.4, 2.0, and 2.2 s were tested. The selection of the subset was based on the results presented in Figure 5.5 from where 9 pairs that are well distributed along the similarity axis (the abscissa) were chosen. The selected piano pairs were: pair 12, 15, 24, 27, 35, 36, 45, 47, and 67. The pairs 47, 35, and 67 were taken from the most similar end of Figure 5.5. The pairs 24, and 27 were taken from the least similar end of the scale. The remaining pairs 12, 15, 36, and 45 were taken from the intermediate similarity range. Then, the simulations for the remaining 13 piano pairs were run using the obtained t_{obs} period. Six threshold estimates were obtained for each piano pair per test condition.

Simulations using the whole dataset of piano sounds

The simulation of discrimination thresholds $\text{thres}_{\text{sim}}$ for the whole dataset of piano sounds (21 piano pairs) was run using the optimal observation period t_{obs} obtained from the exploratory simulations. To further evaluate these simulations, in addition to the comparison of $\text{thres}_{\text{sim}}$ values with Euclidean distances $d_{ij\ \text{exp}}$, simulations of the triadic comparisons using the PEMO model were run. The same simulation scheme for the triadic comparisons as described in Chapter 4 was used for this purpose.

Simulation of triadic comparisons

During a trial, three reverberant piano sounds (A, B, C) were evaluated. For the evaluation their internal representations considering the optimal t_{obs} duration were used. No noise is used because the experimental triadic comparisons were conducted in silence. One CCV value for each of the three possible pairs (AB, AC, and BC) was obtained and, as source of internal variability, a Gaussian noise $N_x(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma = 10.1$ MU[7] is used to obtain $\widehat{\text{CCV}}_{AB}$, $\widehat{\text{CCV}}_{AC}$, and $\widehat{\text{CCV}}_{BC}$ (see Equation 4.8 on page 70). The pair having the maximum $\widehat{\text{CCV}}$ value

[7] Refer to Appendix D (Section D.3.3) for details about the internal noise configuration.

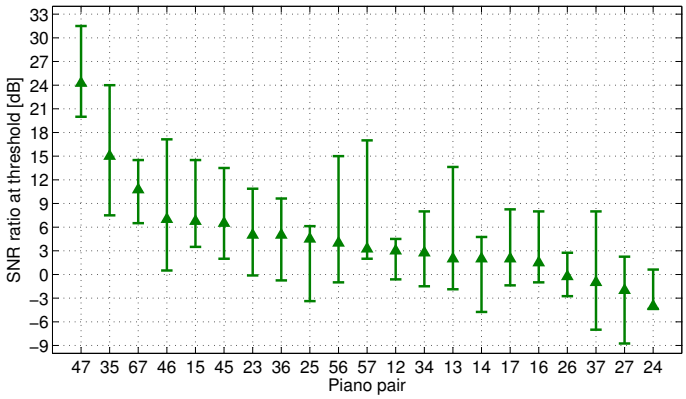


Figure 5.5: Discrimination thresholds for the reverberant piano sounds obtained from the instrument-in-noise tests. The thresholds (green triangles) were assessed taking the median across participants. The piano pairs are shown along the abscissa and are ordered from higher to lower SNR thresholds. The error bars represent interquartile ranges.

was indicated by the artificial listener as the most similar pair. The pair having the minimum $\widehat{\text{CCV}}$ value is indicated as the least similar pair and, therefore, the remaining pair was indicated as having intermediate similarity. To simulate the triadic comparisons of 20 participants, the 35 triads were evaluated 20 times by the artificial listener.

Simulations using ICRA noises version A

The simulations using ICRA noises version A were run for the whole dataset of pianos but using only the duration t_{obs} with the best fit to the experimental data. This choice allows a direct analysis of the spectral differences between the two types of ICRA noises, given that they have similar temporal characteristics and that the artificial listener has access to information during the same observation period t_{obs}.

5.3 Results

5.3.1 Instrument-in-noise test

The discrimination thresholds of the instrument-in-noise experiment are shown in Figure 5.5. The pooled thresholds were assessed by taking the median of all individual threshold estimates per piano pair. The thresholds range between $\text{thres}_{\text{exp,max}} = 24.25$ dB (pair 47) and $\text{thres}_{\text{exp,min}} = -4.0$ dB (pair 24), having a dynamic range $\text{DR}_{\text{exp}} = 28.25$ dB. The estimates have a large between-subject variability with a length of the IQRs from 16.6 dB (pair 46) down to 5.0 dB (pair 24) with a median value of 11.0 dB. The results are based on 189 staircase threshold estimates.

Table 5.3: The similarity matrix S_{ij} derived from the responses of 20 participants (S01-S20) is shown in the upper right triangle. The maximum possible score is 200. The lower left triangle corresponds to the Euclidean distances between stimuli in the resulting four-dimensional space. A high score in the similarity matrix should correspond to a short Euclidean distance. The lowest and highest scores were obtained for the pairs 24 ($S_{ij} = 28$) and 47 ($S_{ij} = 183$). The corresponding distances were 0.93 and 0.19, respectively.

	Piano						
Piano	P1	P2	P3	P4	P5	P6	P7
P1	-	58	78	120	89	110	126
P2	0.85	-	169	28	88	54	48
P3	0.78	0.21	-	78	115	67	67
P4	0.64	0.93	0.79	-	124	119	183
P5	0.76	0.78	0.65	0.61	-	117	118
P6	0.68	0.85	0.84	0.65	0.65	-	144
P7	0.61	0.89	0.79	0.19	0.65	0.50	-

During the data collection 210 staircases were obtained. Twenty-one of the 210 threshold estimates were excluded.

Exclusion criteria

Twenty-one staircases were excluded from the data analysis after the data collection. Seven staircases were removed because the participants reached a maximum SNR of 50 dB ("minimum" noise level). This value was set in advance as floor condition. The remaining 14 thresholds were removed after a check of consistency of the staircases. For this the standard deviation of the reversals was assessed. Thresholds estimations where the deviation of the reversals was larger than 4 dB were removed. It should be noted that this criterion is less strict than the criterion used in Chapter 3, which was based on a standard deviation of 3 dB. If this criterion would have been adopted, 24 other staircases should be excluded (total of 45 exclusions, representing 21% of the data). We decided to keep the criterion of 4 dB to preserve more experimental data points[8].

5.3.2 Triadic comparisons

Construction of the similarity matrix

The results of all participants were pulled out to construct the similarity matrix S_{ij} shown in the upper right triangle of Table 5.3. The matrix was constructed attributing the same similarity counts as in Chapter 3.

Multidimensional scaling

The experimental data were further processed by first converting the similarity scores S_{ij} into counts of dissimilarity D_{ij} (see Equation 3.1). The

[8] Although not shown here, the overall simulated thresholds do not change significantly by adopting either exclusion criterion (3 or 4 dB). The simulated thresholds $\text{thres}_{\text{sim,3 dB}}$ and $\text{thres}_{\text{sim,4 dB}}$ have correlations of $r_p(19) = 0.97$, $p < 0.001$ and $r_s(19) = 0.93$, $p < 0.001$.

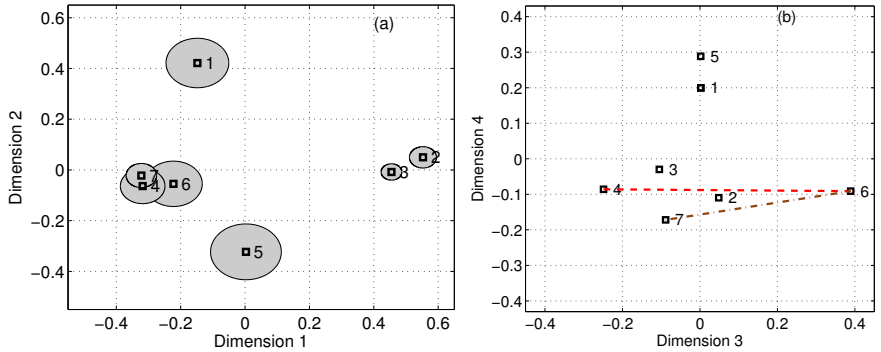


Figure 5.6: Perceptual space obtained with the non-metric MDS algorithm. The dimensions 1 and 2 are shown in panel (a) and the dimensions 3 and 4 in panel (b). This space suggests that the reverberant piano sounds (note C#$_5$) can be grouped in five areas: pianos P2+P3, P4+P7, P1, P5, and P6. Although from the representation of dimensions 1 and 2, piano P6 seems to be similar to P4 and P7, they are far apart along dimension 3 (panel (b)). The stress for the space with dimensions 1 and 2 is poor ($S_t = 29.2\%$). By adding dimension 3 the stress decreases to fair ($S_t = 12.7\%$) and with dimension 4 to nearly good ($S_t = 6.9\%$). The relative distribution of the pianos in the space is not changed in the four dimensional space. The grey bubbles in panel (a) give an indication of the participant's variability. Note that the axes of the MDS spaces are not to scale.

dissimilarity matrix was then used as input for the non-metric MDS algorithm available in the MATLAB Statistics toolbox. An a priori number of $q = 4$ dimensions was used to obtain the perceptual space.

The resulting four-dimensional space has a stress S_t of 6.9% (close to "good"), with cumulative stresses of 29.2% for the first two dimensions ("poor") and 12.7% for the first three dimensions (close to "fair"). The Euclidean distances of the fitted four-dimensional space are shown in the lower left triangle of Table 5.3. The first two dimensions ($S_t = 29.2\%$) of the fitted perceptual space are shown in panel (a) of Figure 5.6. Although this reduced representation provides a poor fit ($S_t > 20\%$), with the exception of piano P6, the overall distribution of the piano sounds in the four-dimensional space is not changed. The relative position of piano P6 gets farther apart from the pianos 47 along dimension 3, as shown in panel (b) of the figure, where the distance d_{46} is 0.64 (red dashed line) and d_{67} is 0.47 (brown dot-dashed line).

The results shown in Figure 5.6 suggest that the reverberant piano sounds can be classified into five distinct groups: pianos P2+P3, P4+P7, P1, P5, and P6. We labelled piano P6 as having intermediate similarity with P4 and P7 despite their overlapped position in the two-dimensional representation of panel (a). This is due to the relative change of the location of P6 when adding the third dimension of the space.

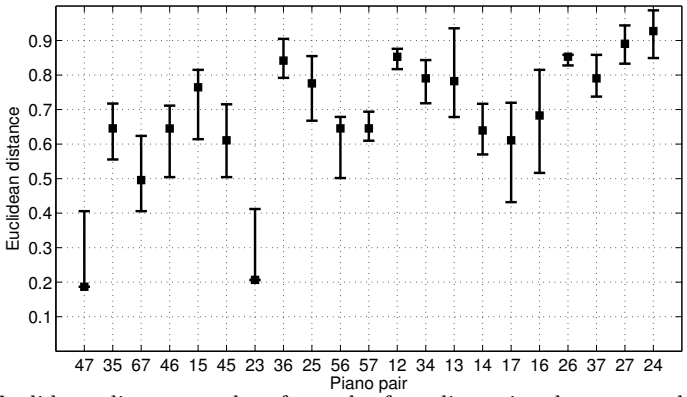


Figure 5.7: Euclidean distances taken from the four-dimensional perceptual space. These distances are also shown in the lower left triangle of Table 5.3. The piano pairs are sorted in the same way as in Figure 5.5. The error bars indicate the minimum and maximum distances between piano pairs across the 5 MDS spaces assessed with data subsets every 4 participants.

Table 5.4: Results of the simulations using a subset of 9 (reverberant) piano pairs and different t_{obs} durations. The minimum and maximum simulated thresholds are indicated together with their dynamic range (DR=$\text{thres}_{\text{max}}-\text{thres}_{\text{min}}$). The correlation values of the simulations with the corresponding experimental data (taken from Figure 5.5) are given. The simulated thresholds of pair 47 were excluded in the assessment of r_p. The SNR range of the experimental data is indicated in column Exp.

		"Observation" duration t_{obs} [s]									
	Exp.	0.16	0.2	0.25	0.3	0.5	0.8	1.0	1.4	2.0	2.2
$\text{thres}_{\text{max}}$ [dB]	24.25	20.75	15.5	13.5	5.25	4.5	0.75	2.75	-0.5	-2.75	-5.5
$\text{thres}_{\text{min}}$ [dB]	-4.0	-4.0	-5.75	-5.5	-5.5	-5.25	-6.5	-7.5	-7.0	-8.5	-9.75
DR [dB]	28.25	24.75	21.25	19.0	10.75	9.75	7.25	10.25	6.5	5.75	4.25
$r_p(6)$	–	0.73*	0.88*	0.74*	0.76*	0.60**	0.42	0.25	0.19	0.25	0.28
$r_s(7)$	–	0.63**	0.80*	0.57	0.75*	0.76*	0.36	0.19	-0.02	0.24	0.37

(*) Significant correlation, $p < 0.05$. (**) Correlations that approach significance, $p < 0.10$.

Between-subject variability

The non-metric MDS algorithm does not provide information about the variability across participants in the resulting fitted space. To inspect individual differences the same approach as described in Chapter 3 was adopted. Five dissimilarity matrices were generated by pulling out the experimental data in groups of 4 participants (S01-S04, S05-S08, S09-S12, S13-S16, and S17-S20). The MDS algorithm was applied to obtain 5 four-dimensional spaces. For each piano sound, the Euclidean distances between these 5 new coordinates and the coordinate in the global space were assessed. Half of the difference between the minimum and the maximum distance is used as radius of the corresponding "bubble" in Figure 5.6. The diameter of the bubbles has a median of 0.14, ranging

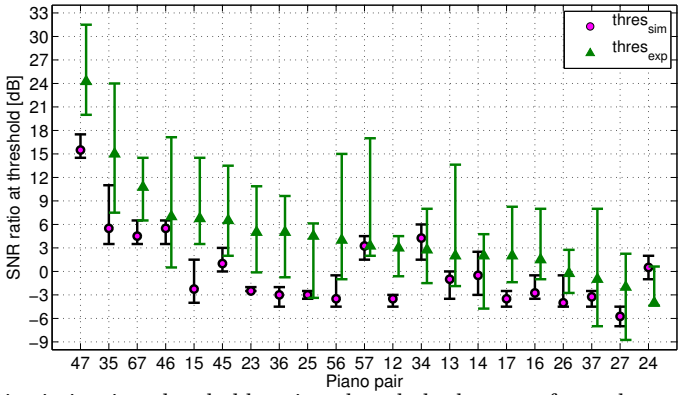


Figure 5.8: Discrimination thresholds using the whole dataset of reverberant piano sounds (21 piano pairs). The median simulated thresholds $\text{thres}_{\text{sim}}$ (for $t_{\text{obs}} = 0.20$ s) are indicated by the magenta circle markers. The green triangle markers correspond to the experimental thresholds $\text{thres}_{\text{exp}}$ (taken from Figure 5.5). The thresholds are shown together with their IQRs. The piano pairs along the abscissa are ordered from higher to lower SNR thresholds based on the experimental data. The thresholds $\text{thres}_{\text{exp}}$ and $\text{thres}_{\text{sim}}$ are significantly correlated with $r_p(18) = 0.58$, $p < 0.01$ and $r_s(19) = 0.61$, $p < 0.001$.

from 0.06 (piano P3) to 0.22 (piano P5). This means that piano P3 was more consistently judged across participants while piano P5 was scored more variable. The obtained 5 four-dimensional spaces were also used to assess the minimum and maximum distances between piano pairs and they are shown as error bars in Figure 5.7. Those deviations ranged between 0.03 (pair 26) and 0.30 (pair 16), with a median length of 0.18.

5.3.3 Instrument-in-noise test: Exploratory simulations

The simulation results of each piano pair are compared with the corresponding experimental thresholds taken from Figure 5.5. The results for the selection of 9 piano pairs are shown in Table 5.4. In the table, information about the minimum and maximum estimated thresholds is shown. Their difference is indicated as DR in dB. As observed in the previous chapter, the simulations that used whole-duration sounds delivered thresholds that are too low with respect to the experimental data. This is visible in the last column of the table, where the results using 2.2-s long piano sounds (whole duration of the sounds plus 0.2 s of silence) delivered thresholds $\text{thres}_{\text{sim}}$ between $\text{thres}_{\text{max}} = -5.5$ dB and $\text{thres}_{\text{min}} = -9.75$ dB, with a DR of 4.25 dB. In order to reduce the information available to the "artificial listener", shorter sections of the piano sounds were fed into the auditory model. The test "observation" durations t_{obs} ranged from 0.16 to 2.2 s. In general, shorter t_{obs} values lead to higher DRs. The only exception was found for $t_{\text{obs}} = 1.0$ s that had a higher DR of 10.25 dB

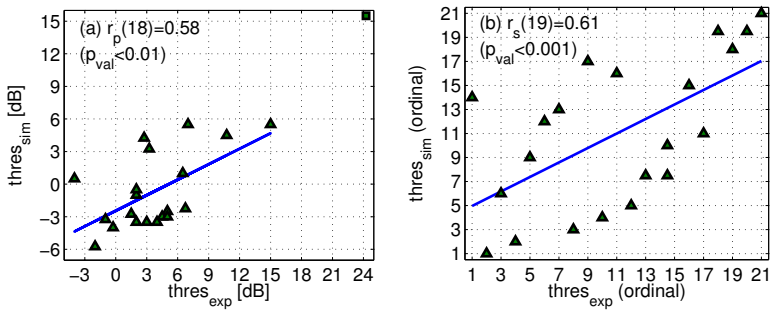


Figure 5.9: Scatter plots and regression analysis between experimental and simulated thresholds of the instrument-in-noise test. The linear regression of panel (a) is related to the Pearson correlation r_p, while the regression in panel (b) to the Spearman (rank-order) correlation r_s. The obtained correlation values were $r_p(18) = 0.58$, $p < 0.01$ and $r_s(19) = 0.61$, $p < 0.001$.

in comparison to the neighbouring t_{obs} durations. The best fit between experimental and simulated data was found for $t_{\text{obs}} = 0.20$ s. For this duration, the thresholds have a Pearson correlation $r_p(6) = 0.88$, $p = 0.01$, and a Spearman (rank-order) correlation $r_s(7) = 0.80$, $p < 0.01$. This "observation" duration was further used to simulate the discrimination thresholds of the remaining 13 piano pairs.

5.3.4 Simulations using the whole dataset of piano sounds

The discrimination thresholds using the whole dataset of piano sounds (21 pairs) were simulated using the first 0.20 s of waveforms (i.e., initial 0.10 s of the piano sounds), based on the results of the exploratory simulations. The median thresholds $\text{thres}_{\text{sim}}$ are indicated by the magenta markers of Figure 5.8. The thresholds are shown together with their IQRs. The simulations at this duration ($t_{\text{obs}} = 0.20$ s) are significantly correlated with the experimental data and a lower but comparable DR_{sim} of 21.25 dB is obtained ($\text{DR}_{\text{sim}} < \text{DR}_{\text{exp}} = 28.25$ dB). The $\text{thres}_{\text{sim}}$ values range between $\text{thres}_{\text{sim,max}} = 15.5$ dB (pair 47) and $\text{thres}_{\text{sim,min}} = -5.75$ dB (pair 27). The Spearman (rank-order) correlation between the thresholds $\text{thres}_{\text{sim}}$ and $\text{thres}_{\text{exp}}$ is significant with $r_s(19) = 0.61$, $p < 0.001$. Although a higher Pearson correlation $r_p(19) = 0.80$, $p < 0.001$ was found, one piano pair (pair 47) was excluded from the regression analysis due to the poor scattering of the $\text{thres}_{\text{exp}}$ values for SNRs above 15 dB[9]. The Pearson correlation after this exclusion is $r_p(18) = 0.58$, $p < 0.01$. The scatter plot of the data together with the corresponding regression analyses are shown in Figure 5.9.

[9]The poor scattering of the data shown in Figure 5.9(a) is also related to the violation of the normality assumption of both, experimental and simulated thresholds.

Table 5.5: Similarity matrix S_{ij} and Euclidean distances derived from the artificial listener using the reverberant piano sounds. The similarity matrix is shown in the upper right triangle and the Euclidean distances between pianos in the resulting four-dimensional space are shown in the lower left triangle. To obtain these results, each triad was evaluated 20. This means that the maximum possible score is 200. The lowest score was obtained for pair 27 ($S_{ij} = 21$) and the highest score was obtained for pairs 13 ($S_{ij} = 171$) and 23 ($S_{ij} = 170$). The corresponding distances were 0.95 (pair 27) and 0.35 (pairs 13 and 23).

	Piano						
Piano	P1	P2	P3	P4	P5	P6	P7
P1	-	124	171	80	87	132	79
P2	0.60	-	170	93	87	94	21
P3	0.35	0.35	-	98	136	110	56
P4	0.76	0.71	0.70	-	66	70	133
P5	0.74	0.74	0.55	0.81	-	77	94
P6	0.57	0.70	0.66	0.81	0.78	-	122
P7	0.77	0.95	0.83	0.55	0.71	0.60	-

Simulation of triadic comparisons

The results of the triadic comparisons using the artificial listener are shown in Table 5.5. In the table, the upper right triangle corresponds to the similarity matrix. A four-dimensional space was obtained using the MDS algorithm. The Euclidean distances between pairs in the fitted space are shown in the lower left triangle of Table 5.5. The obtained space has an excellent goodness of fit (stress $S_t = 2.6\%$) with respect to the similarity matrix. Its first three dimensions (fair stress $S_t = 13.2\%$) are shown in Figure 5.10. The Euclidean distances $d_{ij\ \mathrm{sim}}$ have moderate to weak correlation with $r_p(19) = 0.45$, $p = 0.04$, and $r_s(19) = 0.13$, $p = 0.58$ with respect to the distances $d_{ij\ \mathrm{exp}}$. If stress is used as a measure of correspondence between $d_{ij\ \mathrm{exp}}$ and $d_{ij\ \mathrm{sim}}$ a value $S_{t\ \text{exp-sim}}$ of 25.5% is obtained. Although this value denotes a poor correspondence, it is comparable to the stress $S_{t\ \text{exp-sim}}$ value ($S_{t\ \text{exp-sim}} = 25.2\%$) found for anechoic pianos in Chapter 4. By correlating the dimensions between the "experimental" and "simulated" MDS spaces, the first, second, third, and fourth dimensions have values of $r_s(5) = 0.96, 0.29, 0.54$, and 0.36, respectively. This means that "dimension 1" is the most similarly judged dimension between participants and the artificial listener, followed by the third dimension. The second and fourth dimensions are weighted differently in both MDS spaces.

Simulations using ICRA noises, version A

The discrimination thresholds $\mathrm{thres}_{\mathrm{sim,A}}$ using the whole dataset of piano sounds were simulated using the obtained t_{obs} of 0.20 s. The median

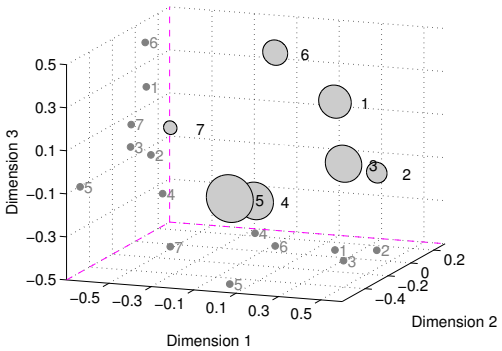


Figure 5.10: Perceptual space obtained from simulated triadic comparisons and with MDS. The first three (of four) dimensions are shown. The grey bubbles give an indication of the "participant's" variability: the bigger the bubble the higher the variability across participants. For ease of visualisation, the location of each piano sound is projected onto dimensions 1-2 (bottom plane) and 2-3 (left plane). Note that the axes of this MDS space are not to scale.

thresholds $\text{thres}_{\text{sim,A}}$ are indicated by the red square markers of Figure 5.11. The thresholds are shown together with their IQRs. For ease of comparison, the simulated thresholds $\text{thres}_{\text{sim}}$ of Figure 5.8, which used version B of the ICRA noise algorithm, are also indicated in Figure 5.11 using magenta circle markers. The $\text{thres}_{\text{sim,A}}$ values range between $\text{thres}_{\text{sim,A,max}}$= 17.0 dB (pair 47) and $\text{thres}_{\text{sim,A,min}}$= −2.5 dB (pair 26). The Spearman (rank-order) correlation between $\text{thres}_{\text{sim,A}}$ and $\text{thres}_{\text{sim}}$ is significant with $r_s(19) = 0.56$, $p < 0.001$. When excluding one piano pair (pair 47, for similar reasons as earlier in this chapter), a significant Pearson correlation of $r_p(18) = 0.61$, $p < 0.01$ is obtained. The scatter plots between $\text{thres}_{\text{sim}}$ and $\text{thres}_{\text{sim,A}}$ are shown in Figure 5.12.

5.3.5 Euclidean distances and instrument-in-noise thresholds

The Euclidean distances obtained from the experimental triadic comparisons $d_{ij\ \text{exp}}$ (from Figure 5.7 and also in the lower left triangle of Table 5.3) and the experimental instrument-in-noise thresholds $\text{thres}_{\text{exp}}$ (from Figure 5.8) have correlation values of $r_p(18) = -0.49$, $p = 0.03$, and $r_s(19) = -0.65$, $p = 0.001$. The corresponding regression analyses and scatter plots are shown in Figure 5.13. In turn, the $d_{ij\ \text{exp}}$ distances and the simulated thresholds $\text{thres}_{\text{sim}}$ have correlation values of $r_p(18) = -0.26$, $p = 0.27$, and $r_s(19) = -0.49$, $p = 0.03$. The correlation values between Euclidean distances obtained from simulated triadic comparisons $d_{ij\ \text{sim}}$ and $\text{thres}_{\text{exp}}$ thresholds are $r_p(19) = -0.34$, $p = 0.14$, and $r_s(19) = -0.27$, $p = 0.23$, and with $\text{thres}_{\text{sim}}$ thresholds

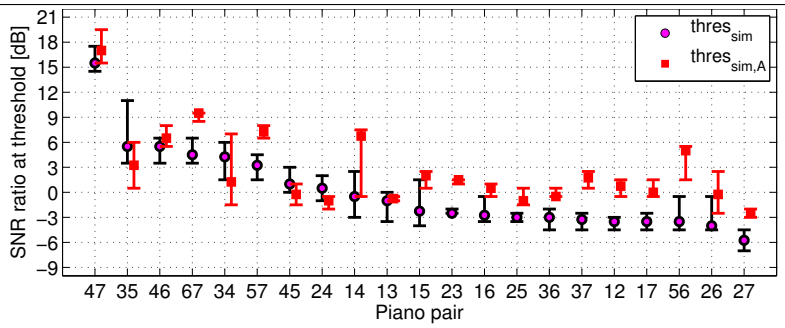


Figure 5.11: Simulated thresholds using the whole dataset of reverberant piano sounds ($t_{\text{obs}} = 0.20$ s) and different types of ICRA noise. The median simulated thresholds $\text{thres}_{\text{sim}}$ using ICRA noises version B are indicated by the magenta circle markers (same as in Figure 5.8). The red square markers correspond to simulated thresholds $\text{thres}_{\text{sim,A}}$ obtained using ICRA noises version A. The thresholds are shown together with their IQRs. The piano pairs along the abscissa are ordered from higher to lower SNR thresholds based on $\text{thres}_{\text{sim}}$. The $\text{thres}_{\text{sim}}$ and $\text{thres}_{\text{sim,A}}$ are significantly correlated with $r_p(18) = 0.61$, $p < 0.01$ and $r_s(19) = 0.59$, $p < 0.001$.

are $r_p(19) = -0.23$, $p = 0.31$, and $r_s(19) = -0.31$, $p = 0.17$. Finally, we report the correlation between $d_{ij\ \text{sim}}$ distances and $\text{thres}_{\text{sim,A}}$ thresholds obtained using ICRA noises version A. Their correlation values are $r_p(19) = -0.14$, $p = 0.53$, and $r_s(19) = -0.11$, $p = 0.65$. These values show that $d_{ij\ \text{sim}}$ distances and $\text{thres}_{\text{sim,A}}$ thresholds are not correlated.

All correlation values reported in this chapter are summarised in the schema of Figure 5.14.

5.4 Discussion

5.4.1 Comparison between experimental methods

A high perceptual similarity is equivalent to a high SNR threshold and a short Euclidean distance. Scatter plots between the median thresholds $\text{thres}_{\text{exp}}$ from the instrument-in-noise test (taken from Figure 5.5) and the Euclidean distances from the triadic comparisons (taken from Figure 5.7) were shown in Figure 5.13 together with corresponding linear regression analyses. The thresholds $\text{thres}_{\text{exp}}$ were found to be significantly correlated with the Euclidean distances ($r_p(18) = -0.49$, $p = 0.03$, and $r_s(19) = -0.65$, $p = 0.001$). The median thresholds $\text{thres}_{\text{exp}}$ have an IQR of 4.7 dB ($\text{thres}_{25-75} = 1.9 - 6.6$ dB). The Euclidean distances have an IQR of 0.17 ($d_{ij,25-75} = 0.63 - 0.80$).

Further inspection of the data shown in Figures 5.5 and 5.7 reveals that both methods share 2 of the 3 most similar pairs (pairs 47 and 36). The methods also coincide in the judgement of 3 out of the 5

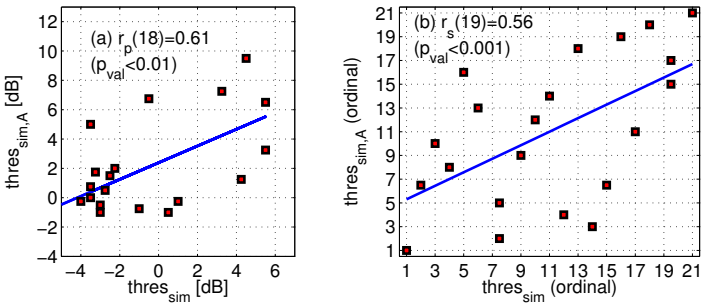


Figure 5.12: Scatter plots and regression analysis between simulated thresholds using different types of ICRA noise. The linear regression of panel (a) is related to the Pearson correlation r_p, while the regression in panel (b) to the Spearman (rank-order) correlation r_s. One pair of points was removed from the analysis to obtain an $r_p(18) = 0.61$, $p < 0.01$, due to the lack of $\text{thres}_{\text{sim}}$ values above 6 dB. A Spearman correlation of $r_s(19) = 0.59$, $p < 0.001$ is obtained.

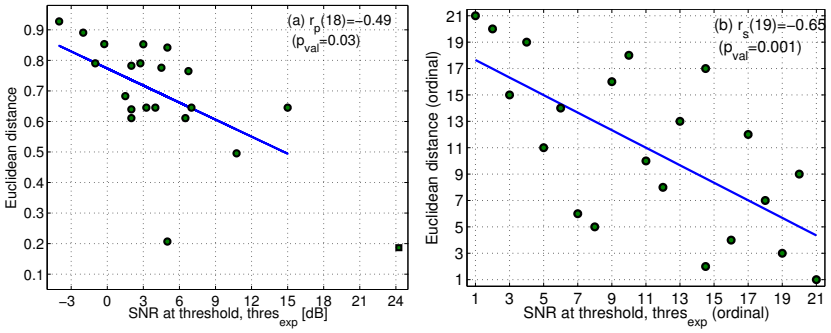


Figure 5.13: Scatter plots and regression analysis between the results of the instrument-in-noise (SNR thresholds, $\text{thres}_{\text{exp}}$) and triadic comparisons tests (Euclidean distances). The linear regression of panel (a) is related to the Pearson correlation r_p, while the regression in panel (b) to the Spearman (rank-order) correlation r_s. One pair of points was removed from the analysis to obtain an $r_p(18) = -0.49$, $p = 0.03$, due to the lack of $\text{thres}_{\text{exp}}$ values above 15 dB. A Spearman correlation of $r_s(19) = -0.65$, $p = 0.001$.

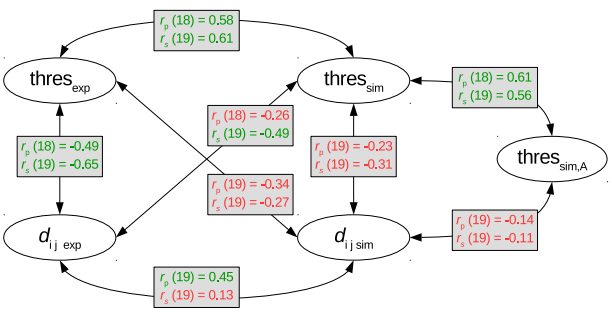


Figure 5.14: Summary of correlation values between instrument-in-noise thresholds and Euclidean distances. All possible combinations among $\text{thres}_{\text{exp}}$, $\text{thres}_{\text{sim}}$, $d_{ij\ \text{exp}}$, and $d_{ij\ \text{exp}}$ are indicated in this schema. The correlation values of the simulated thresholds using the ICRA noise algorithm version A with $\text{thres}_{\text{sim}}$ and $d_{ij\ \text{sim}}$ are also indicated.

most different pairs ($\text{thres}_{\text{exp}} < 1.9$ dB and distance > 0.80): 24, 26, 27. There are, however, some pairs for which the methods provide different similarity measures. If the IQR of the thresholds and distances are used to delimit three similarity regions: high ($d_{ij,25} \leq 0.63$, $\text{thres}_{\text{exp},75} \geq 6.6$ dB), medium (d_{ij}, $\text{thres}_{\text{exp}}$ within their IQRs), and low similarity ($d_{ij,75} \geq 0.80$, $\text{thres}_{\text{exp},25} \leq 1.9$ dB), five piano pairs are judged differently by the two methods. These pairs are:

- Pair 15: the distance $d_{15} = 0.76$ indicate that pianos P1 and P5 are more distinct than the threshold $\text{thres}_{\text{exp},15}$ indicates.
- Pair 36: the distance $d_{36} = 0.84$ indicate that pianos P3 and P6 are more distinct than the threshold $\text{thres}_{\text{exp},36}$ indicates.
- Pair 12: the distance $d_{12} = 0.85$ indicate that pianos P1 and P2 are more distinct than the threshold $\text{thres}_{\text{exp},12}$ indicates.
- Pair 23: the distance $d_{23} = 0.21$ indicate that pianos P2 and P3 are more similar than the threshold $\text{thres}_{\text{exp},23}$ indicates.
- Pair 16: the distance $d_{16} = 0.68$ indicate that pianos P1 and P6 are more distinct than the threshold $\text{thres}_{\text{exp},16}$ indicates.

The higher number of discrepancies in the judgement of both methods may be related to the apparent increase of difficulty of the task with respect to the comparison between anechoic pianos of Chapter 3. Evidence for this are: (1) the lower stress values of the fitted MDS space with respect to the experimental similarity matrix $S_{t\ \text{rev}} = 6.9\%$ in contrast to $S_{t\ \text{ane}} = 3.1\%$ (from Chapter 3), and the poorer cumulated stress for the first two and three dimensions ($S_{t\ \text{rev}} = 29.2$ and 12.7% compared with $S_{t\ \text{ane}} = 21.9$ and 7.5%, respectively) (2) the larger number of excluded staircases in case the same criterion as in Chapter 3 would have been adopted in the current chapter. Despite the discrepant judgements and the apparent increase in the difficulty of the tasks, the rank-order correlation between methods ($r_s(19) = -0.65$, $p = 0.001$) is statistically not different from the value obtained in Chapter 3 ($r_s(19) = -0.64$, $p = 0.001$, see panel (b) of Figure 3.9).

5.4.2 Comparison between experimental and simulated thresholds

The simulated thresholds $\text{thres}_{\text{sim}}$ of the instrument-in-noise method are significantly correlated with the experimental thresholds $\text{thres}_{\text{exp}}$

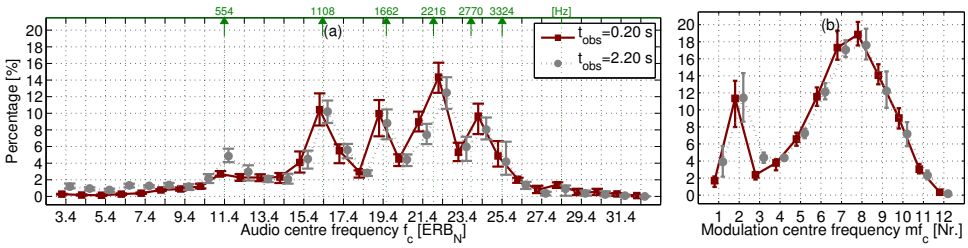


Figure 5.15: Weighting of information in difference (internal) representations ($\Delta R_x \cdot T_p$) for whole-duration sounds (grey circle markers) and considering a shorter duration t_{obs} of 0.20 s (maroon square markers). The weighting I_m/I_{tot} of each audio frequency channel is shown in panel (a). The weighting I_k/I_{tot} of each modulation frequency channel is shown in panel (b). The values per band are expressed as percentage with respect to the total information I_{tot} of each representation. The points along the ERB scale that correspond to F0= 554 Hz and its five first harmonics are indicated by the green labels on the top axis.

($r_p(18) = 0.58$, $p < 0.01$ and $r_s(19) = 0.61$, $p < 0.001$) when only the initial part of the waveforms is used. A duration $t_{\text{obs}} = 0.20$ s provided the best fit between $\text{thres}_{\text{sim}}$ and $\text{thres}_{\text{exp}}$. This is in line with the simulation results of Chapter 4, where a t_{obs} of 0.25 s was used. Scatter plots between the median $\text{thres}_{\text{sim}}$ and $\text{thres}_{\text{exp}}$ thresholds (taken from Figure 5.8) are shown in Figure 5.9 together with corresponding linear regression analyses. The discussion in this subsection is based on an analysis of the information that is integrated by the artificial listener to produce the obtained $\text{thres}_{\text{sim}}$ values. This analysis is, in turn, based on the processing of information in the template-weighted piano representations ($\Delta R_x \cdot T$) per audio (I_m/I_{tot}) and modulation frequency channel (I_k/I_{tot}) given by Equations 4.6 and 4.7, as used in the previous chapter. The contribution of each frequency band (I_m/I_{tot} or I_k/I_{tot}) was assessed using the total duration of the piano-plus-noise sounds (2.2 s) and using only the first 0.20 s of the waveforms. In this analysis, all 21 pairs and their corresponding ICRA noises were used. The noises were set to the level indicated by the corresponding simulated threshold $\text{thres}_{\text{sim}}$. The information-weighted values together with their IQRs are shown in Figure 5.15. The weighting I_m/I_{tot} of each audio frequency channel is shown in panel (a) of the figure. The weighting I_k/I_{tot} of each modulation frequency channel is shown in panel (b) of the figure. The band weightings for both t_{obs} durations are very similar overall with mean differences $\Delta I_m/I_{\text{tot}}$ of 0.0% (IQR= 1.08%) and $\Delta I_k/I_{\text{tot}}$ of 0.0% (IQR= 1.64%).

For the information in the audio frequency channels I_m/I_{tot} (panel (a) of Figure 5.15), most of the information is comprised in bands around

the first five harmonics ($15.4 < f_c < 25.4$ ERB_N) with 80.6% of the information for representations with $t_{\text{obs}} = 0.20$ s and 74.5% for representations with $t_{\text{obs}} = 2.20$ s. At the band centred at the F0 of the piano note ($f_c = 11.4$ ERB_N) the representations with $t_{\text{obs}} = 2.20$ s provide a slightly higher (but still low) weighting of 4.9% in comparison with the 2.7% given by the representations with $t_{\text{obs}} = 0.20$ s. For both durations the weighting of information at the F0 band is lower than the weighting found for the anechoic piano sounds (see panel (a) of Figure 4.10, maroon markers) that had a cumulative weighting of about 10% in the bands centred at 11 and 12 ERB_N.

For the information in the modulation frequency channels I_k/I_{tot} (panel (b) of Figure 5.15), the filters Nr. 2 and 6−9 have an individual weighting of 10% or more, comprising 73.1% and 70.4% of the information in the representations that use a t_{obs} of 0.2 and 2.2 s, respectively. In comparison with the weighting of information for anechoic piano sounds (see panel (b) of Figure 4.10, maroon markers), the second modulation filter ($\text{mf}_c = 5$ Hz) has a lower value of 11.3% ($t_{\text{obs}} = 0.2$ s) which is 7.3% less than the value of 18.6% in the anechoic piano representations ($t_{\text{obs}} = 0.25$ s). For higher modulation filters, especially for bands 6 ($\text{mf}_c = 46.3$ Hz) to 9 ($\text{mf}_c = 214.3$ Hz), the weighting of information has become more prominent, reaching a values 17.7 and 18.8% at bands 7 ($\text{mf}_c = 77.2$ Hz) and 8 ($\text{mf}_c = 128.6$ Hz), respectively. These values are about 3% higher than the values for the anechoic piano representations. The changes in the weighted information per modulation filter may be attributed to the reverberation applied to the piano sounds, that introduces more variations or cues in the colour of the piano sounds. This reduces the relative importance of the envelope information, which is conveyed mainly in the first three modulation filters.

The (overall) similar weighting for the reverberant sounds using either observation period t_{obs} (shorter or longer duration) may lead us to the same hypothesis of Chapter 4 about the prominent role of the internal noise in the success of the simulated $\text{thres}_{\text{sim}}$ values. We have confirmed this hypothesis and, although the results are not shown here, the analysis presented in Section 4.6.2 is also applicable for these reverberant sounds.

5.4.3 Comparison between simulated thresholds for different ICRA noise versions

The simulated thresholds $\text{thres}_{\text{sim}}$ of the instrument-in-noise method using ICRA noises version B are significantly correlated with the thresh-

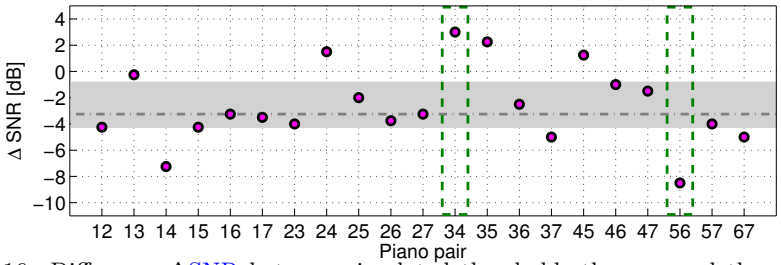


Figure 5.16: Difference ΔSNR between simulated thresholds $\text{thres}_{\text{sim}}$ and $\text{thres}_{\text{sim,A}}$ (see Figure 5.11), that were obtained using ICRA noises version B (as in this chapter) and A, respectively. A ΔSNR value below 0 dB indicates that the SNR threshold $\text{thres}_{\text{sim,A}}$ is higher than the $\text{thres}_{\text{sim}}$ for the corresponding piano pair. The shadowed area indicates the IQR of the ΔSNR values and the median of −3.25 dB is indicated by the horizontal (grey) dashed line. The ΔSNR values for pairs 34 and 56 are further analysed in the text.

olds $\text{thres}_{\text{sim,A}}$ obtained using ICRA noises version A ($r_p(18) = 0.61$, $p < 0.01$ and $r_s(19) = 0.56$, $p < 0.001$). Both sets of thresholds were obtained using representations limited to $t_{\text{obs}} = 0.2$ s. The simulation results were shown in Figure 5.11 and their corresponding regression analyses in Figure 5.12. The difference ΔSNR between simulated thresholds ($\text{thres}_{\text{sim}} - \text{thres}_{\text{sim,A}}$) is shown in Figure 5.16. The median difference ΔSNR across all piano pairs is −3.25 dB (indicated by the horizontal grey dashed-dotted line in the figure) with an IQR between −4.25 dB and −0.8 dB. This means that on average, ICRA noises version A produce discrimination thresholds ($\text{thres}_{\text{sim,A}}$) that have a higher SNR (i.e., a lower noise level) than the thresholds ($\text{thres}_{\text{sim}}$) obtained using ICRA noises version B. Based on the IQR of ΔSNR values (shadowed area in Figure 5.16), we may classify the piano pairs into three groups:

1) Pairs with SNR thresholds that are above percentile 75 ($\text{thres}_{\text{sim}} - \text{thres}_{\text{sim,A}} > -0.8$ dB): pairs 13, 24, 34, 35, and 45.

2) Pairs with SNR thresholds that are within the IQR ($-4.25 \leq \text{thres}_{\text{sim}} - \text{thres}_{\text{sim,A}} \leq -0.8$ dB): pairs 12, 15, 16, 17, 23, 25, 26, 27, 36, 46, 47, and 57.

3) Pairs with SNR thresholds that are below percentile 25 ($\text{thres}_{\text{sim}} - \text{thres}_{\text{sim,A}} < -4.25$ dB):pairs 14, 37, 56, and 67.

To further evaluate the ΔSNR differences we take the two piano pairs that have the maximum and minimum ΔSNR value: pair 34 (ΔSNR = 3 dB, from "Group 1") and pair 56 (ΔSNR = −8.5 dB, from "Group 3").

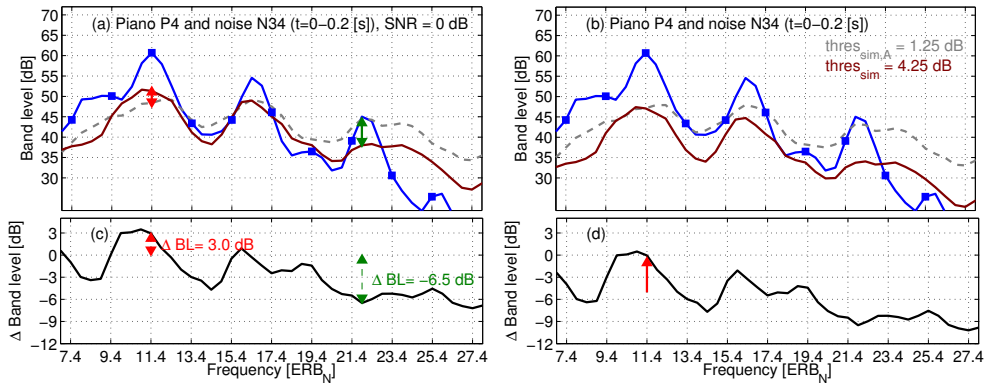


Figure 5.17: Band levels for piano P4 (blue solid line) and paired noise N34 using ICRA noises version A (grey dashed line) and B (maroon solid line) at an SNR of 0 dB (panel (a)) and at their simulated threshold (panel(b)). In the bottom panels the difference in band levels between ICRA noises version A and B are shown. The red and green arrows in panels (a) and (c) indicate the frequencies 11.4 ERB_N and 21.9 ERB_N, respectively, where the absolute difference in band levels is greater than 3 dB. As shown in panels (b) and (d), the band level of the ICRA noises at 11.4 ERB_N are approximately the same when the noises A and B are plotted at their simulated thresholds $thres_{sim,A}$ of 1.25 dB and $thres_{sim}$ of 4.25 dB. This is indicated by a Δ Band level of 0 dB in panel (d), indicated by the red arrow.

For better understanding the subsequent analysis it is important to bear in mind the effect of using either ICRA algorithm on the noise band levels with respect to the band levels of the corresponding piano sounds. As pointed out earlier in this chapter, the ICRA noises used to obtain $thres_{sim}$ (version B) and $thres_{sim,A}$ (version A) have the same overall level with respect to the corresponding piano sounds, but the ICRA noises version A have a spectral tilt with increasing band levels towards higher frequencies. This is a relative increase in level that reaches a level difference of 10 dB in the highest auditory band with respect to the F0-centred filter. This means for the two noises that for a given ΔSNR, there should be one spectral band for which, after compensating for the threshold difference, the band levels of the two noises are equal.

For the analysis of pair 34, band levels of three signals –piano P4[10] and paired ICRA noise N34 in versions A and B– are shown in panel (a) of Figure 5.17. In panel (c) the difference in band levels ΔBL between the two noise versions is shown. The red and green arrows indicate points in frequency where the absolute difference $\|\Delta BL\|$ is greater than 3 dB. Hence, those differences may have produced the non-zero

[10]The choice of using pianos P4 and P6 in the analyses shown in Figures 5.17 and 5.18 is based on the fact that the leading criterion used by the artificial listener for the selected pairs 34 and 56 is, in both cases, criterion 2 (i.e., using the template $T_{p,r}$ derived from the reference piano). In these pairs the reference pianos are P4 and P6, respectively.

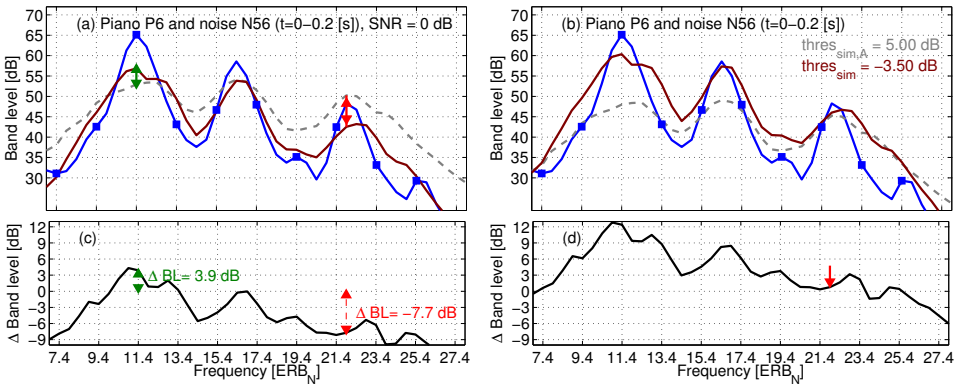


Figure 5.18: Band levels for piano P6 (blue solid line) and paired noise N56 using ICRA noises version A (grey dashed line) and B (maroon solid line) at an SNR of 0 dB (panel (a)) and at their simulated threshold (panel(b)). In the bottom panels the difference in band levels between ICRA noises version A and B are shown. The green and red arrows in panels (a) and (c) indicate the frequencies 11.4 ERB_N and 21.9 ERB_N, respectively, where the absolute difference in band levels is greater than 3 dB. As shown in panels (b) and (d), the band level of the ICRA noises at 21.9 ERB_N are approximately the same when the noises A and B are plotted at their simulated thresholds $\mathrm{thres}_{\mathrm{sim,A}}$ of 5 dB and $\mathrm{thres}_{\mathrm{sim}}$ of −3.5 dB.

ΔSNR indicated by the green dashed rectangles in Figure 5.16. To further investigate which frequency region does actually produce the difference between $\mathrm{thres}_{\mathrm{sim}}$ and $\mathrm{thres}_{\mathrm{sim,A}}$ values, the band levels for the same sounds are replotted in panel (b) of Figure 5.17, but using the SNRs at threshold for the paired noises ($\mathrm{thres}_{\mathrm{sim,A}}$= 1.25 dB for version A and $\mathrm{thres}_{\mathrm{sim}}$= 4.25 dB for version B). The difference indicated by the red arrow in panel (c) where ΔBL is 3.0 dB at 11.4 ERB_N (higher band level for noise version B) seems to have been equated for the noises at threshold shown in panel (d) that have a ΔBL of 0 dB. A similar analysis can be applied to pair 56 (piano P6, noise N56 versions A and B). The analysis is shown in Figure 5.18. The difference ΔBL= −7.7 dB at 21.9 ERB_N (higher band level for noise version A) indicated by the red arrow in panel (c) is reduced to ΔBL= 0.78 dB ≈ 0 dB when the level of the noises at the simulated thresholds is used (panel (d), $\mathrm{thres}_{\mathrm{sim,A}}$= 5 dB for version A and $\mathrm{thres}_{\mathrm{sim}}$= −3.5 dB for version B).

The previous analyses provided evidence that for pair 34 (from "Group 1") the most relevant audio frequencies used by the artificial listener lie around 11.4 ERB_N (near the F0 of the note) and for pair 56 (from "Group 3") around 22.9 ERB_N (near the partial at $f = 4{\cdot}\mathrm{F0} = 2216$ Hz). All other piano pairs have ΔSNR*s* between the values for pairs 34 and 56 (see Figure 5.16). The (ICRA) noise band with equal level at the

corresponding thresholds $\text{thres}_{\text{sim,A}}$ and $\text{thres}_{\text{sim}}$ will therefore be in the spectral range between harmonics F0 and 4·F0 of the C#$_5$ note.

If the efficiency of the two noises is to be evaluated in terms of the amount of noise needed to mask the properties of the piano sounds, then ICRA noises version A perform "better" because for the same overall (broad-band) noise level the discrimination thresholds have on average higher SNRs (lower noise level) compared to ICRA noises version B: ΔSNR = −3.25 dB (IQR between −4.25 and −0.75 dB). This "better performance" is, however, at the expense of a gradual level mismatch towards higher frequencies of the noises with respect to the sounds to be masked. If the efficiency of the noises is to be evaluated in terms of how well do the spectro-temporal properties of the noise follow the properties of the sounds to be masked then ICRA noises version B perform better.

5.5 Conclusion

In this chapter the instrument-in-noise method of Chapters 3 and 4 has been applied to the same dataset of pianos to which the effect of reverberation was added by digital convolution. Experimental thresholds $\text{thres}_{\text{exp}}$ were collected using a new version of the ICRA noise algorithm and compared with Euclidean distances obtained from experimental triadic comparisons $d_{ij\ \text{exp}}$. The results of both methods had a similar correlation compared to the values reported in Chapter 3, with a Pearson correlation $r_p(18) = -0.49$, $p = 0.03$ and a Spearman correlation $r_s(19) = -0.65$, $p = 0.001$. Using the same simulation scheme as in Chapter 4, estimates $\text{thres}_{\text{sim}}$ of the instrument-in-noise method were obtained using the PEMO model. In order to bring the $\text{thres}_{\text{sim}}$ thresholds to the range of $\text{thres}_{\text{exp}}$, the observation period of the artificial listener had to be reduced to $t_{\text{obs}} = 0.20$ s. The obtained $\text{thres}_{\text{sim}}$ values had correlations of $r_p(18) = 0.58$, $p < 0.01$, and $r_s(19) = 0.61$, $p < 0.001$.

An information-based analysis of the internal representations obtained from the PEMO model showed that the effect of a 3-s long reverberation on our set of piano notes (C#$_5$) increased the importance of the audio frequency bands (I_m/I_{tot}) comprising the first five harmonics above the F0 (between 15.4 and 25.4 ERB_N) and decreased the importance of the band centred at the F0 of the note with respect to the weightings found for the anechoic pianos in Chapter 4. In terms of the information conveyed by the modulation filters (I_k/I_{tot}), the filters $6-9$ ($\text{mf}_c = 46.3-128.6$ Hz) increased their relative weighting while the lower

modulation filters ($\text{mf}_c \leq 10$ Hz) decreased their importance.

Further simulations were used to address the following aspects: (1) the estimation of discrimination thresholds $\text{thres}_{\text{sim,A}}$ using ICRA noises version A, and (2) the simulation of the triadic comparison task. The estimated $\text{thres}_{\text{sim,A}}$ and $\text{thres}_{\text{sim}}$ values had correlations $r_p(18) = 0.61$, $p < 0.01$, and $r_s(19) = 0.56$, $p < 0.001$. For the first aspect, an analysis based on the difference between thresholds indicated that the decisions of the artificial listener are importantly influenced by the information contained in the spectral region between F0 and $4 \cdot$ F0. For the second aspect, we had a limited success with the simulation of the triadic comparisons, where the simulated distances $d_{ij\ \text{sim}}$ had only medium to low correlations with both experimental and simulated thresholds. An analysis of the resulting MDS spaces revealed that only their first and third dimensions were correlated with high or moderate values $r_p(5) = 0.96$ and $r(5) = 0.54$, respectively. It is important to note, however, that the same simulation approach reached only a moderate correlation with $d_{ij\ \text{exp}}$ in Chapter 3. Moreover we found some evidence for an increase in the task difficulty with respect to the experiments using anechoic pianos (Chapter 3). In the instrument-in-noise method, for instance, the consistency in the staircases adopting the exclusion criterion of Chapter 3 would have led to more exclusions: 38 staircases (18.1% of the data) in contrast to the 24 staircases (11.4%) excluded using a more permissible criterion. Additionally, the goodness of fit of the MDS space (from the experimental sessions) had a somewhat poorer fit with respect to the collected similarity matrix. This may be due to a more variable weighting of psychological dimensions for different participants.

In summary, the experimental results presented in this chapter showed that instrument-in-noise thresholds are similarly correlated with Euclidean distances from the triadic comparisons for the perceptual similarity assessment of reverberant piano sounds with respect to the results reported in Chapter 3 for anechoic sounds. Furthermore, simulations of the instrument-in-noise thresholds using the PEMO model had a similar degree of success with respect to the simulations of Chapter 4. We can conclude that the results of this chapter further support the validity of the auditory modelling approach of Chapter 4 when the effect of reverberation is applied to the dataset of sounds.

6 Simulating the perceived reverberation in different room acoustic environments using a binaural auditory model[1]

In this chapter an alternative way to use the unified modelling framework introduced in Chapter 4 is presented. Particularly, a binaural model is used to compute estimates of perceived reverberation –known as reverberance– for a set of musical instrument sounds that are auralised using eight different acoustic environments. The binaural model processes left and right-ear channels in a similar manner as the auditory PEMO model used in the previous chapters, but the central processor converts the (left- and right-ear) internal representations into a metric of reverberance P_{REV}. This central processor is based on the idea of stream segregation, adopted from the field of auditory scene analysis, rather than on the optimal detector used in previous chapters to approach the problem of perceptual similarity among sounds.

In the first part of the chapter, P_{REV} estimates obtained from the binaural auditory model, originally described and validated by van Dorp (2011) and van Dorp, de Vries, and Lindau (2013), are compared with the room acoustic parameters of reverberation time (T_{30}) and early decay time (EDT). For this comparison, 90-s music excerpts of an orchestra consisting of 23 instrument sections are used. The simulation results show that although P_{REV} has a higher correlation with EDT than with T_{30}, this relationship depends on the properties of the instruments. Further analyses show that P_{REV} depends on the presentation level and that for instruments with similar critical-band spectrum, P_{REV} follows a similar trend across acoustic conditions.

In order to obtain experimental evidence of the dependency of reverberance on the properties of the sounds being tested, a listening test

[1] This chapter is partly based on: A. Osses, A. Kohlrausch, W. Lachenmayr, and E. Mommertz (2017). Predicting the perceived reverberation in different room acoustic environments using a binaural model. *J. Acoust. Soc. Am.*, 141(4), EL381-EL387. http://doi.org/10.1121/1.4979853

is presented in the second part of this chapter. The stimuli used in the listening test correspond to a subset of 8 musical instrument sounds that had been used in the initial simulations. The experimental results support the hypothesis that the sensation of reverberance is instrument-dependent. Furthermore, these results are used to evaluate the validity of the P_{REV} estimates obtained from the binaural model.

6.1 Introduction

A set of binaural room impulse responses (BRIRs) is usually measured in order to evaluate the acoustic characteristics of a room. Out of these impulse responses conventional descriptors such as reverberation time (RT) or, in this chapter, T_{30}, EDT and clarity index (C_{80}) are obtained. To obtain those parameters, the guidelines established in the international standard ISO 3382-1 (ISO, 2009) can be followed. This guarantees a reproducibility of the measurement results. The measurements are often performed in empty rooms. Since the acoustic descriptors do differ when measured in empty or occupied halls (see, e.g., Beranek, 2004) the latter condition is always of interest, especially in the context of a concert hall or an opera house. Partly motivated by this idea, van Dorp et al. (2013) suggested the use of a time-domain binaural auditory model to estimate room acoustic parameters. Their rationale was that sound samples recorded or simulated in a given acoustic environment convey room acoustic cues that can be extracted by a binaural auditory model. A similar assumption was made by Klockgether and van de Par (2014, 2016) who used excerpts of violin, guitar, and snare drum sounds to investigate the spatial attributes of listener envelopment (LEV) and apparent source width (ASW), and the JND in the binaural cues of interaural level difference (ILD) and interaural time difference (ITD) in three acoustic environments.

During the development of their binaural model, van Dorp et al. (2013) conducted four listening experiments using two sounds (speech and cello), which were auralised using 27 BRIRs. They found that their model estimates were highly correlated with the subjective percept of reverberation, known as reverberance.

Motivated by the success of their model, in this chapter we present an extension of their work by analysing a more diverse set of sounds in acoustic conditions that are typical for rehearsal and music performance venues. Our set of sounds consisted of 23 instruments from a 90-s ex-

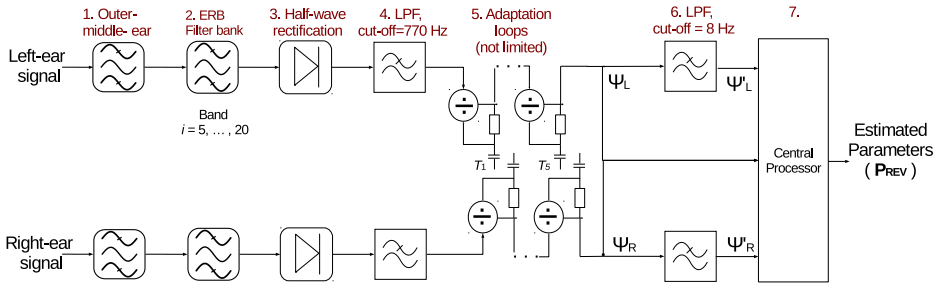


Figure 6.1: Block diagram of the binaural auditory model. The stages 1 to 7 are briefly described in the text. A parallel processing of the left and right ear signals is followed by a central processing stage, where both "internal representations" (Ψ'_{L}, Ψ'_{R}) are combined to obtain the model estimates.

cerpt of an orchestra recording that are individually analysed using the binaural auditory model. In order to account for the long duration of the sound samples, a frame-based approach was followed to obtain reverberance estimates as a function of time. To allow this and other slight modifications of the model, we implemented the binaural model using the framework of the AMT toolbox for MATLAB (Søndergaard & Majdak, 2013), introducing the central processor as described by van Dorp et al. (2013).

6.2 The binaural auditory model

The binaural auditory model used in this chapter is referred to as Room Acoustic Analyser (RAA) and is described in detail by van Dorp (2011). The block diagram of the model is shown in Figure 6.1. The RAA model is based on the model described by Breebaart et al. (2001) but implementing an alternative central processor (Stage 7 in the figure). The model is applied separately to left and right-ear signals followed by a central processor. The monaural stages of the model are:

Stage 1. Outer- and middle-ear filtering: This stage is implemented as a second-order bandpass IIR filter between 1000 and 4000 Hz. This implementation corresponds to a simpler approximation to the actual filtering introduced by outer and middle ear compared to the implementation shown in Chapter 4. The combined frequency response of the outer and middle ear is shown in Figure 6.2.

Stage 2. Gammatone filter bank: This set of filters corresponds to an approximation to a critical-band filter bank. The filter bank consists of

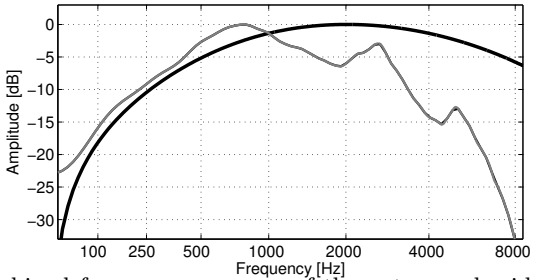


Figure 6.2: The combined frequency response of the outer- and middle-ear filters as used in the binaural model is indicated by the black thick line. The second-order BPF used in the model is a simpler implementation with respect to the filters used in the PEMO model, whose frequency response is indicated by the grey line (see also Figure 4.2).

16 bands having centre frequencies between 165 (5 ERB_N[2]) and 1750 Hz (20 ERB_N). The Gammatone filter bank is implemented in the same way as described in Chapter 4.

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 low-pass filtered ($f_{\text{cut-off}} = 770$ Hz). These stages are implemented in the same way as described in Chapter 4.

Stage 5. Adaptation loops: This stage simulates the adaptive properties of the auditory periphery and it differs from the description given in Chapter 4 in two parameters: (1) One of the short time constants was replaced by a longer one ($\tau_1 = 5$ ms, $\tau_2 = 129$ ms, $\tau_3 = 253$ ms, $\tau_4 = 376$ ms, and $\tau_5 = 500$ ms), and (2) no overshoot limitation is applied, i.e., the limiter factor for the RAA model tends to infinity ($\mathsf{limit} \rightarrow \infty$). This configuration was also used by Breebaart et al. (2001) and van Dorp (2011) and in earlier versions of the monaural auditory models.

Stage 6. Modulation low-pass filter: In this stage the signal (internal) representations are smoothed by means of a single-pole LPF with a time constant of 20 ms ($f_{\text{cut-off}} = 8$ Hz). This stage is used instead of the modulation filter bank in the PEMO model. The modulation low-pass filter provides a similar smoothing as that introduced by the lowest modulation filter of the PEMO model but they differ in their cut-off frequencies.

[2]The ERB rate scale corresponds to one of the frequency scales that is inspired by the frequency representation in the auditory system. A brief overview of this scale is given in Appendix A.

6.2.1 Central processor

To couple both monaural outputs, a central processor is used (stage 7 in Figure 6.1). The incoming signals are segregated into a "foreground" stream Ψ'_{dir} and a "background" stream Ψ'_{rev}. These streams are assumed to be related to the direct sound coming from the sound source and the acoustic environment in which the sound source is embedded, respectively. Within each auditory band k, an algorithm is used to detect peaks with durations longer than T_{min} above the threshold $\Psi_{\text{min}}(k)$. The detection is also used to detect dips longer than T_{min} with values below the threshold $\Psi_{\text{min,dip}}(k)$. These threshold values are proportional to the average band level $L_{\Psi}(k)$:

$$\begin{aligned} \Psi_{\text{min}}(k) &= \mu_{\Psi} \cdot L_{\Psi}(k) \\ \Psi_{\text{min,dip}}(k) &= \mu_{\Psi,\text{dip}} \cdot L_{\Psi}(k) \end{aligned} \tag{6.1}$$

To obtain the average level in the kth band $L_{\Psi}(k)$, the absolute value of the amplitudes $\Psi'[n,k]$ (after stage 6 in Figure 6.1) are arithmetically averaged in time.

As a result of the peak detection algorithm, the N-sample streams Ψ'_L (and Ψ'_R) are classified into $\Psi'_{\text{L,dir}}$ (and $\Psi'_{\text{R,dir}}$) or $\Psi'_{\text{L,rev}}$ (and $\Psi'_{\text{R,rev}}$). Next, left (L) and right (R) channels are combined. For the amplitudes of the background stream:

$$\Psi_{\text{rev}}[n,k] = \sqrt{\left(\Psi'_{\text{L,rev}}[n,k]\right)^2 + \left(\Psi'_{\text{R,rev}}[n,k]\right)^2} \tag{6.2}$$

Finally, by arithmetically averaging the levels Ψ_{rev}, a total reverberance level L_{rev} is obtained:

$$P_{\text{REV}} = L_{\text{rev}} = \frac{1}{N \cdot K} \sum_{n=0}^{N-1} \sum_{k=k_0}^{k_1} \Psi_{\text{rev}}[n,k] \tag{6.3}$$

where K is the total number of frequency bands being used ($K = k_1 - k_0 + 1$). The values for the constants used in Equations 6.2 and 6.3 are shown in Table 6.1. As indicated in Equation 6.3, the reverberant level L_{rev} is used as reverberance estimate P_{REV} and it is expressed in MU.

Although P_{REV} is only based on the reverberance level L_{rev}, the average level L_{dir} can be similarly obtained using Equations 6.2 and 6.3:

$$L_{\text{dir}} = \frac{1}{N \cdot K} \sum_{n=0}^{N-1} \sum_{k=k_0}^{k_1} \Psi_{\text{dir}}[n,k] \tag{6.4}$$

Table 6.1: Parameters of the RAA model as reported by van Dorp ("Original") and as used in our implementation ("Our") for estimating P_{REV}.

Parameter	Values: Our (Original)	Description
$k_0 - k_1$	5-20 (5-20)	Initial and final spectral band number (ERB_N) used in the estimation ($K = 16$ bands)
f_c [Hz]	174-1807 (168-1836)	Centre frequencies of the initial and final ERB band used in the estimation
μ_Ψ	0.34 ($7.49 \cdot 10^{-3}$)	Constant factor for peak detection
$\mu_{\Psi,\mathrm{dip}}$	-0.06 ($-1.33 \cdot 10^{-3}$)	Constant factor for dip detection
$\lvert\mu_\Psi/\mu_{\Psi,\mathrm{dip}}\rvert$	5.63 (5.63)	Ratio between peak detection factors
T_{min} [ms]	63.1 (63.1)	Minimum peak/dip duration for the foreground stream

The level L_{dir} is used by the central processor of the model to obtain three other room acoustics estimates which are not described in this thesis (van Dorp, 2011; van Dorp et al., 2013).

6.2.2 Differences in the current implementation

Our implementation of the binaural model differs slightly from the original RAA model. We did not account for the absolute threshold of hearing, originally implemented as a frequency-dependent scaling before and after stage 5 (adaptation loops). As a consequence of this, the amplitudes of the internal representations differ (leading to different band levels L_Ψ) and, therefore, different μ-factors were required to get an appropriate segregation of the foreground and background streams. The parameters used in our implementation are shown in Table 6.1. In order to deal with sounds containing silent sections, only those segments where each instrument was active were considered.

6.3 Study case: Reverberance of different orchestra instruments

6.3.1 Rooms

Four rooms have been simulated in the software Odeon Auditorium v.13 using the suggested accuracy "engineering". Three of the rooms were simulated with different absorptions, producing a total number of 8 acoustic environments (i.e., 8 "rooms"). Some information about the 8 acoustic environments is given in Table 6.2. The acoustic parameters were estimated at the location of a binaural listener arbitrarily placed 7 m in front of the stage in all cases. The room A is a medium-sized music venue with a coupled ceiling space, simulated without (A) and with absorption on the walls and the ceiling ($\mathsf{A}_{\mathsf{abs}}$). The room B is a large-sized concert hall, simulated without (B) and with all interior walls

Table 6.2: List of rooms used in this chapter. The EDT and T_{30} values were obtained as an arithmetic average of 23 estimations (obtained from the available BRIRs in each room) at mid frequencies (500-1000 Hz). The column G* gives an indication of the sound strength in the rooms. For ease of interpretation of the results in the subsequent sections, the rooms are sorted by increasing EDT times.

Room / Description of the hall	Volume [m^3]	Seats	EDT [s]	T_{30} [s]	G* [dB]
A_{abs} / Medium-sized, coupled space (abs. 720 m^2)	14000	1000	0.80	1.14	5.4
B_{abs} / Large-sized (abs. 3700 m^2)	23000	2600	0.81	1.20	0.0
A / Medium-sized, coupled space	14000	1000	0.83	1.51	7.3
C_{abs1}/ Rehearsal (abs. 250 m^2)	2500	100	1.04	1.16	9.8
B / Large-sized	23000	2600	1.24	2.01	1.3
C_{abs2}/ Rehearsal (abs. 190 m^2)	2500	100	1.27	1.34	10.8
D / Medium-sized	15000	1300	1.47	2.23	8.5
C / Rehearsal	2500	100	2.48	2.51	12.7

(*) The sound strength G is a measure of relative energy with respect to an impulse response measured at a distance of 10 m. In this study, however, we first computed the integrated sound pressure level per instrument, and then those 23 levels were arithmetically averaged in each room. The softest averaged level was used as a 0 dB reference (room $\mathsf{B}_{\mathsf{abs}}$). Therefore the assessed G* values indicate how much louder a room is with respect to the reference room.

absorptive ($\mathsf{B}_{\mathsf{abs}}$). The room C is an orchestra rehearsal space modelled in three conditions: with 250 m^2 ($\mathsf{C}_{\mathsf{abs1}}$) and 190 m^2 ($\mathsf{C}_{\mathsf{abs2}}$) of absorption and without any acoustic treatment (C). The room D corresponds to a medium-sized concert hall (Fog & Ballinger, 2008). All rooms were set as occupied ($\alpha_\omega = 0.9$) in the simulations, with the exception of room C where no additional audience (only musicians) was considered.

Considering a JND in RT[3] of about 0.1 s, the rooms $\mathsf{A}_{\mathsf{abs}}$ and $\mathsf{B}_{\mathsf{abs}}$ do not differ by more than one JND and neither they do with respect to room A if only EDT is considered. Likewise, rooms B and $\mathsf{C}_{\mathsf{abs2}}$ do not significantly differ from each other when considering the averaged EDT values. A difference of less than one JND means that the respective rooms cannot be distinguished based on their reverberation time.

6.3.2 Stimuli

The sounds consist of 23 anechoic recordings of orchestra instruments[4] that were used as sound sources in the Odeon software to simulate a medium-sized orchestra of 56 musicians (some recordings were used more than once), divided into four sections:

[3]The JND for EDT is a relative value of 5% (ISO, 2009). For our minimum and maximum EDT times the JND is 0.04 s and 0.13 s, respectively.

[4]The sounds were derived from anechoic symphony orchestra recordings (Rindel, 2015) made at the Technical University of Denmark (DTU) and licensed to Odeon A/S. The WavePackInstall II containing the anechoic recordings can be obtained at http://www.odeon.dk/anechoic-recordings.

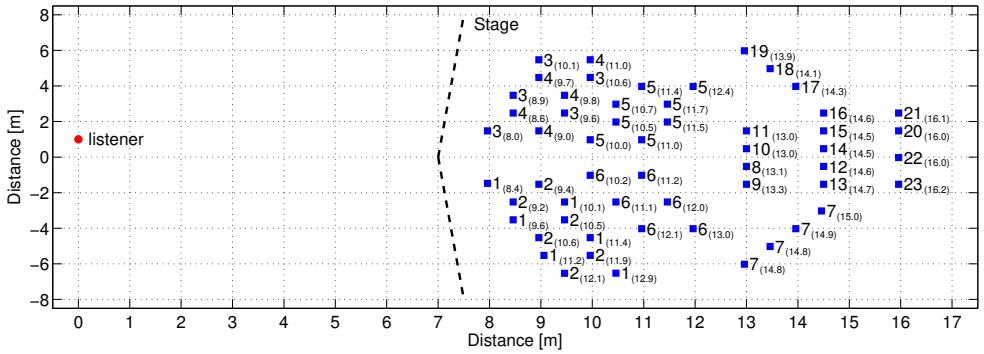


Figure 6.3: Distribution of the orchestra as used in Odeon for room D. The distances along the abscissa are referenced to the position of the virtual listener. The listener is located 7 m in front of the stage and 8 m far away from the closest "musician". This position is located in the audience area already which is further extended behind the listener (it would correspond to negative distances in this figure, not shown). The musicians are indicated by numbers between 1 and 23* (see Table 6.3 for the corresponding labels). The numbers between brackets represent the (Euclidean) distance in m from each musician to the listener.

(*) Note that the location of the first (Nr. 1-2) and second violins (Nr. 3-4) to the right and left of the audience area (virtual listener), respectively, does not match a typical orchestra distribution. It would be more natural to have the first violins to the left and the second violins to the right. The distribution shown in this figure is, however, the configuration as used in the (existing) Odeon project to which we had access to.

• **Strings (40 musicians)**: first violin (Nr. 1, x 6), first violin retake (Nr. 2, x 6), second violin (Nr. 3, x 5), second violin retake (Nr. 4, x 5), viola (Nr. 5, x 8), cello (Nr. 6, x 6), double bass (Nr. 7, x 4);

• **Woodwind (9 musicians)**: flute (Nr. 8, x 1), piccolo (Nr. 9, x 1), oboe (Nr. 10-11, x 2), clarinet (Nr. 12-13, x 2), bassoon (Nr. 14-15, x 2) and contrabassoon (Nr. 16, x 1);

• **Brass (5 musicians)**: French horn (Nr. 17-19, x 3), trumpet (Nr. 20-21, x 2), and;

• **Percussion (2 musicians)**: timpani (Nr. 22, x 1), triangle (Nr. 23, x 1).

The instruments were distributed on the available stage area as similar as possible in each venue. A virtual listener 7 m in front of the stage was added, leading to average listener-musician distances between 9.7 to 16.8 m (min-max distances of 7.8-18.6 m). The distribution of instruments ("musicians") on the stage area of room D is shown in Figure 6.3.

Auralisation

The auralisations were automatically generated in Odeon. For this process, static directivity patterns were used for each instrument, obtaining

Table 6.3: Instruments available in the orchestra. The distances to the binaural listener (7 m in front of the stage) and the sound levels of the auralised sounds were averaged across rooms. The levels $L_{\mathrm{Aeq},T}$ (A-weighted) and $L_{\mathrm{Zeq},T}$ (linear) were integrated over the duration T and their difference is indicated as ΔL_{eq}.

Nr./ Instrument	Distance [m] to listener	Sound levels [dB]				T	ΔL_{eq}
		$L_{\mathrm{Aeq},T}$	L_{AFmax}	$L_{\mathrm{Zeq},T}$	L_{ZFmax}	[s]	[dB]
1-4/ Violins (Vio)	10.6 (8.0-12.9)	68.0	77.1	68.1	80.4	357	0.1
5/ Viola (Viola)	10.2 (7.8-12.4)	71.7	80.9	73.9	87.3	88	2.2
6/ Cello (Cello)	9.7 (7.8-14.8)	66.6	74.9	74.9	86.8	87	8.3
7/ Double bass (DBass)	12.4 (10.4-15.0)	65.4	73.8	84.8	97.1	77	19.4
8/ Flute (Flute)	13.4 (13.1-13.9)	79.7	88.7	78.9	90.5	45	-0.8
9/ Piccolo (Picc)	13.3 (13.0-13.8)	66.4	74.2	65.4	76.9	45	-1.0
10-11/ Oboe (Oboe)	13.8 (13.0-15.4)	74.0	81.7	73.7	85.1	107	-0.3
12-13/ Clarinet (Cla)	14.8 (14.5-15.4)	70.2	78.0	72.0	81.3	130	1.8
14-15/ Bassoon (Bsn)	14.8 (14.5-15.3)	66.8	73.5	70.3	78.8	139	3.5
16/ Contrabassoon (CBsn)	14.7 (14.2-15.9)	56.2	64.7	64.1	74.3	70	7.9
17-19/ French horn (FrHrn)	14.9 (13.6-17.3)	71.3	78.5	75.8	85.2	141	4.5
20-21/ Trumpet (Trum)	16.2 (15.6-18.5)	76.0	84.4	76.1	86.4	88	0.1
22/ Timpani (Ti)	16.5 (15.7-18.6)	70.6	78.2	84.5	95.5	38	13.9
23/ Triangle (Tri)	16.8 (16.2-18.6)	64.3	73.7	66.8	80.4	28	2.5

Table 6.4: Correlation between the P_{REV} values and EDT and T_{30}.

	Correlation with			Correlation with	
Nr./ Instrument	EDT	T_{30}	Nr./ Instrument	EDT	T_{30}
1-4/ Vio	0.91*	0.82*	12-13/ Cla	0.92*	0.75*
5/ Viola	0.94*	0.84*	14-15/ Bsn	0.97*	0.77*
6/ Cello	0.90*	0.90*	16/ CBsn	0.20	0.55
7/ DBass	0.78*	0.57	17-19/ FrHrn	0.95*	0.73*
8/ Flute	0.51	0.39	20-21/ Trum	0.46	0.10
9/ Picc	0.47	0.37	22/ Ti	0.90*	0.86*
10-11/ Oboe	0.53	0.43	23/ Tri	0.18	-0.23

(*) Significant correlation, $p < 0.05$.

56 different BRIRs at the location of the listener. These BRIRs were internally used by Odeon to auralise anechoic recordings of a 90-s excerpt of the Brahms Symphony Nr. 4, 3rd movement. The auralised strings were mixed down per instrument obtaining 7 waveforms (first violin x 2, second violin x 2, viola x 1, cello x 1, double bass x 1), reducing the total number of auralised sounds from 56 to 23. Hence, Odeon returned 23 BRIRs and 23 binaural sounds. Information about the sound levels of the resulting sounds is shown in Table 6.3.

6.3.3 Using the auditory model

The sounds corresponding to the 23 instruments listed in Table 6.3, auralised in the 8 different acoustic conditions (total of 184 binaural signals)

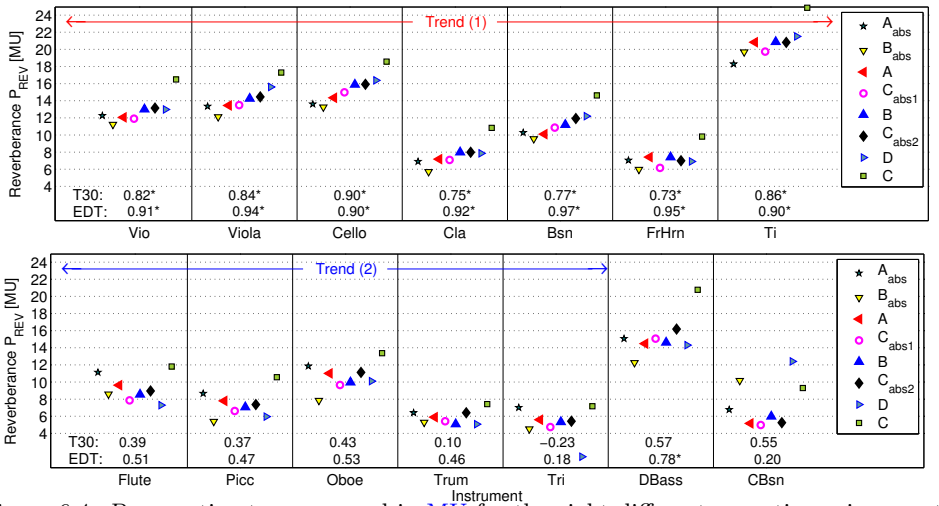


Figure 6.4: P_{REV} estimates expressed in MU for the eight different acoustic environments. For each instrument the correlation between P_{REV} and T_{30} and EDT is shown. Values marked with asterisks indicate that the corresponding P_{REV} estimate is linearly related with T_{30} and/or EDT ($p<0.05$).

were fed into the RAA model. Reverberance estimates (P_{REV} expressed in MU) were obtained for 5-s long sections and 80% overlap, leading to a total of 86 values per sound sample. Subsequently, the estimates from the same instruments were grouped together to finally use the median in each group as single P_{REV} estimate. As a consequence of this, the sounds were reorganised in 14 groups. Within each group, 8 estimates were obtained (one estimate per room).

6.4 Results

The results obtained from the 184 auralised instrument sounds are shown in Figure 6.4. The overall model estimates range from a minimum value of 4.2 MU (CBsn) to a maximum value of 22.4 MU (Ti). Although this represents a variation of 18.2 MU, the difference between estimates within each instrument group (ΔP_{REV}) is smaller and ranges from 1.6 (Trum) to 7.5 MU (DBass) with a median ΔP_{REV} of 4.6 MU, indicating that the P_{REV} estimates are instrument dependent. When analysing the relative P_{REV} values, some trends can be observed: (1) Vio, Viola, Cello, Cla, Bsn, FrHrn and Ti: the lowest P_{REV} is attributed to room B_{abs}, similar estimates are obtained for A, C_{abs1} and also for B, C_{abs2}, D and highest P_{REV} is obtained for room C; (2) Flute, Picc, Oboe, Trum and Tri: the lowest and highest P_{REV} are also attributed to the rooms B_{abs} and C,

Table 6.5: Pearson correlation r_p between the model estimates P_{REV} for all possible instrument pairs. The matrix is symmetric along its diagonal. For instance, the highest correlation ($r_p = 0.98$) for violin estimates (Vio) is obtained for the comparison with the clarinet estimates (Cla). Likewise, the lowest correlation ($r_p = 0.21$) is obtained for the comparison with the contrabassoon estimates (CBsn).

	Instrument													
	Vio	Viola	Cello	DBass	Flute	Picc	Oboe	Cla	Bsn	CBsn	FrHrn	Trum	Ti	Tri
Vio	-	0.94	0.92	0.94	0.57	0.78	0.78	0.98	0.96	0.21	0.94	0.72	0.89	0.34
Viola	0.94	-	0.96	0.84	0.32	0.61	0.71	0.95	0.98	0.34	0.83	0.53	0.88	0.03
Cello	0.92	0.96	-	0.83	0.21	0.53	0.60	0.95	0.97	0.23	0.79	0.45	0.91	0.03
DBass	0.94	0.84	0.83	-	0.66	0.88	0.87	0.94	0.89	-0.03	0.90	0.85	0.78	0.52
Flute	0.57	0.32	0.21	0.66	-	0.91	0.78	0.49	0.35	-0.12	0.72	0.88	0.31	0.86
Picc	0.78	0.61	0.53	0.88	0.91	-	0.95	0.76	0.62	-0.19	0.88	0.90	0.53	0.77
Oboe	0.78	0.71	0.60	0.87	0.78	0.95	-	0.79	0.69	-0.14	0.85	0.86	0.54	0.58
Cla	0.98	0.95	0.95	0.94	0.49	0.76	0.79	-	0.96	0.12	0.93	0.66	0.90	0.30
Bsn	0.96	0.98	0.97	0.89	0.35	0.62	0.69	0.96	-	0.29	0.82	0.59	0.89	0.11
CBsn	0.21	0.34	0.23	-0.03	-0.12	-0.19	-0.14	0.12	0.29	-	0.10	-0.12	0.33	-0.54
FrHrn	0.94	0.83	0.79	0.90	0.72	0.88	0.85	0.93	0.82	0.10	-	0.75	0.84	0.49
Trum	0.72	0.53	0.45	0.85	0.88	0.90	0.86	0.66	0.59	-0.12	0.75	-	0.49	0.74
Ti	0.89	0.88	0.91	0.78	0.31	0.53	0.54	0.90	0.89	0.33	0.84	0.49	-	0.08
Tri	0.34	0.03	0.03	0.52	0.86	0.77	0.58	0.30	0.11	-0.54	0.49	0.74	0.08	-

respectively (with the exception of the Flute), but the remaining rooms, sorted by increasing estimates are D, $\mathrm{C_{abs1}}$, B, $\mathrm{C_{abs2}}$, A and $\mathrm{A_{abs}}$; (3) two other different patterns were observed for DBass and CBsn. For DBass, room D was "judged" as the second least reverberant hall, while for CBsn room C was the third most reverberant room and one inconsistent within-room P_{REV} was found ($\mathrm{B_{abs}}$ had a higher estimate than B). The instruments following the trend (1) had higher correlations with EDT (all significant) than with T_{30} (6 of 7 significant correlations). None of the instruments following trend (2) had a significant correlation with EDT nor T_{30}.

Another way of comparing the reverberance trends is to construct a similarity matrix based on the correlation between the P_{REV} estimates of all possible instrument pairs. Such a matrix is shown in Table 6.5. This matrix can be further processed by techniques as the MDS algorithm (already used in Chapters 3 and 5) to map each of the stimuli (14 groups of instruments) to a graphical Cartesian representation. A two-dimensional representation of the instruments is shown in Figure 6.5. The instruments belonging to trends (1) and (2) are indicated by the black and red square markers, respectively. The DBass and CBsn sounds, which were identified as following different reverberance patterns, are indicated by white markers in the figure.

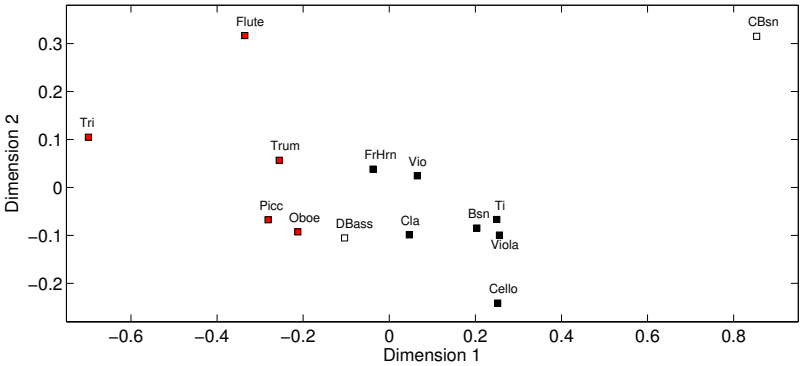


Figure 6.5: Cartesian two-dimensional representation (stress S_t = 22.2%) of the 14 instrument groups of the orchestra based on a similarity of the "reverberance trends". This analysis is based on the correlation matrix shown in Table 6.5. Instruments that are close to each other provide a similar ranking in their P_{REV} values for the eight tested acoustic environments. Three trends were recognised (and described in the text), trend (1), trend (2), and trend (3). They are indicated by the black, red, and white square markers, respectively.

6.5 Interim discussion

The reverberance estimates $\mathrm{P_{REV}}$ obtained from the RAA model were found to be instrument dependent. This is in agreement with the results presented by Teret, Pastore, and Braasch (2017) for three reverberance-matching experiments with 5 types of sounds (orchestra, broad-band noise, click, guitar, and voice samples), where "signal type" was found to be significant. From their set of sounds, the guitar and voice samples were found to be the samples eliciting the lowest and highest reverberance, respectively. In another study, Klockgether and van de Par (2014) also found room acoustics estimates depending on the analysed sound (guitar, violin or snare drum). In order to understand the differences across instruments in our approach, the following aspects are addressed: (a) which properties do the instruments following trends (1) and (2) share; (b) what is the most prominent property influencing the P_{REV} amplitude range, and; (c) how large is the variability in the P_{REV} range within instruments.

6.5.1 Spectral content

Twenty-one of the 23 orchestra sounds (91.3% of the data) had a P_{REV} estimate following either trend (1) (14 sounds, 60.9% of the data) or (2) (7 sounds, 30.4% data). The two remaining instruments (DBass, CBsn) had P_{REV} estimates following other trends (8.7% of the data). Our analysis is therefore focused on these two trends. Since P_{REV} depends on the stream segregation performed in the central processor stage and, in turn, it de-

pends on the average band level within each critical-band, the energy distribution of two representative musical instruments following trends (1) and (2) is shown in panel (a) of Figure 6.6. The instruments following trend (1) had a balanced spectrum with contributions between roughly 5 and 10% per band. The instruments following trend (2) had a monotonic increasing contribution from nearly 0% (Picc) up to around 10% towards the upper bands. Therefore, in order to characterise a room obtaining one single estimate in the whole frequency range, it might be desirable to use instruments following trend (1). Since the auditory filters are narrower in the low frequency range a higher spectral level at low frequencies is needed. An estimate that can give an indication of the frequency distribution is the difference between linear and A-weighted levels. The instruments following trend (1), with the exception of the violins, have a ΔL_{eq} that varies between 1.8 (Cla) and 13.9 dB (Ti) (see Table 6.3)[5].

6.5.2 Frame-based values

As the individual instruments have dynamic changes along their 90 s of music ($6.7 \leq L_{\mathrm{AFmax}} - L_{\mathrm{Aeq,T}} \leq 9.4$ dB, see Table 6.3), we hypothesised that the reverberation estimate should also vary over time. The adopted frame-based approach is useful to provide information about changes of reverberance as a function of time. In Figure 6.6(b) the data points corresponding to rooms $\mathsf{A}_{\mathsf{abs}}$ and B are shown together with bars indicating the minimum and maximum P_{REV} values over time. This variability is systematic in all instruments and the average range is ±3.2 MU.

6.5.3 Level dependency

To investigate the dependency of P_{REV} on presentation level, three of the instrument groups (Vio, Flute, Ti) were plotted at two presentation levels with a level difference of 20 dB. The obtained estimates are shown in panel (b) of Figure 6.6. For the three instrument groups, P_{REV} increased when increasing the presentation level. Evidence of the reverberance dependency on presentation level was given by Lee, Cabrera, and Martens (2012), where louder test samples required bigger adjustments to match their reverberance with respect to a fixed-gain reference sample. Within the RAA model, the increase in the estimates seems to be further related to the instrument spectral properties, with a stronger effect for the Flute (factor of 3) followed by the Ti (factor of 1.6) and the Violin (factor of 1.4).

[5] Although the level estimation shown in Table 6.3 is valid for the auralised sounds in room D, they are representative approximations of the level difference in the other seven acoustic conditions.

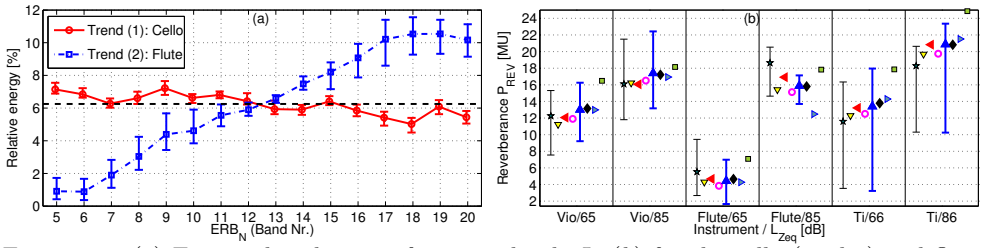


Figure 6.6: (a) Energy distribution of average levels $L_\Psi(k)$ for the cello (circles) and flute samples (squares). The bars indicate the minimum and maximum levels across rooms. For equal band contribution the $L_\Psi(k)$ levels should follow the horizontal dashed line (6.25%). (b) P_{REV} estimates for Vio, Flute and Ti at two presentation levels of either 65 or 66 dB and 20 dB more intense. The markers for each room are the same as in Figure 6.4. The estimates for rooms $\mathsf{A}_{\mathsf{abs}}$ and B are shown together with their minimum and maximum values.

6.6 Listening experiment

So far the binaural RAA model has been used to obtain P_{REV} estimates for a set of recordings auralised in eight different acoustic environments. The results show that P_{REV} depends on the spectral content of the sound being processed and on the presentation level. The RAA model has been previously validated using two samples (voice and cello) in a large number of acoustic environments, but we have no indication of the validity of the model for the set of orchestra sounds used so far. In order to validate our implementation of the RAA model with a selected set of sounds and, in turn, provide evidence that not only the simulated P_{REV} but also experimental $P_{\mathrm{REV,exp}}$ values are instrument-dependent, a listening test designed to evaluate the perceived reverberation is presented in this second part of the chapter.

The experiment was designed in a way that the duration of each experimental session lasts no more than one hour. The aim of the experiment was to sort the sound samples from least to most reverberant. A preference method was adopted, for which the multi-stimulus comparison method (see Section 1.3.3) was preferred to pairwise comparisons due to its time efficiency. However, it was necessary to reduce the number of samples to be evaluated (8 excerpts per trial) and the duration of each sound (10-s long excerpts instead of 90-s long). For this reason, the stimulus treatment differs from what it was done in the P_{REV} simulations of the previous sections.

6.6.1 Stimuli

A subset of the instruments described in Section 6.3.2 was chosen. The stimuli were chosen to be representative of the results obtained from the

Table 6.6: Level information about the instruments of the Odeon orchestra used in the listening experiment. The sound levels of the auralised sounds were averaged across rooms. The levels $L_{\mathrm{Aeq},T}$ (A-weighted) and $L_{\mathrm{Zeq},T}$ (linear) were integrated over the entire duration of 10 s and their difference is indicated as ΔL_{eq}. The column "Δ Pres. level" is obtained as the difference between the maximum value of the auralised waveforms L_{ZFmax} and the maximum of the 90-s sounds used in the simulations (L_{ZFmax} of Table 6.3). All differences are negative, meaning that a softer reproduction level is used in the listening experiments in comparison with the assumed levels in the simulations of Figure 6.4.

Nr./ Instrument	Sound levels [dB]				ΔL_{eq}	Δ Pres.
	$L_{\mathrm{Aeq},T}$	L_{AFmax}	$L_{\mathrm{Zeq},T}$	L_{ZFmax}	[dB]	level [dB]
1/ Vio	65.4	73.1	65.1	73.3	-0.3	-7.1
7/ DBass	62.4	69.7	80.7	88.2	18.3	-8.9
8/ Flute	69.1	78.7	68.4	77.8	-0.8	-12.7
9/ Picc	67.3	76.0	66.3	74.8	-1.1	-2.1
16/ CBsn	49.7	55.8	57.0	62.9	7.2	-11.4
17/ FrHrn	68.8	75.8	72.9	79.4	4.1	-5.8
20/ Trum	73.9	81.0	73.9	80.8	0.0	-5.6
22/ Ti	65.7	73.9	79.3	89.2	13.6	-6.3

simulations (Figure 6.4). In this way, three instruments with a reverberance estimate from trend (1) were chosen: violin (Vio, Nr. 1), French horn (FrHrn, Nr. 17), and timpani (Ti, Nr. 22); three instruments from trend (2): flute (Flute, Nr. 8) piccolo (Picc, Nr. 9), and trumpet (Tr, Nr. 20); and the two instruments that followed "another" trend: double bass (DBass, Nr. 7), and contrabassoon (CBsn, Nr. 16). The subset of instruments consisted thus of 8 instruments. Excerpts of no more than 10 s were chosen. The excerpts were taken from the first 18 bars of the symphony, where most of the instruments play *fortissimo*.

Auralisation

The reverberant orchestra sounds were obtained by digital convolution of the 8 selected anechoic recordings with the corresponding BRIR, which were previously obtained from Odeon. The convolution was performed in MATLAB. A fixed gain of -9 dB was applied to the resulting sounds to prevent clipping after auralisation. The resulting waveforms had levels that we labelled as comfortable. Therefore no further level adjustment was applied. Information about the (average) sound levels of the auralised sounds is shown in Table 6.6.

6.6.2 Apparatus

The experiments were conducted in a single-walled sound booth. The stimuli were presented via Sennheiser HD 265 Linear circumaural headphones in a binaural reproduction. The participant's responses were

collected using the software Web Audio Evaluation (WAE) (Jillings et al., 2016) using Google Chrome on a local computer.

6.6.3 Participants

Twenty-four participants (5 females and 19 males) were recruited from the JF Schouten subject database of the TU/e university. At the time of testing, the participants were between 19 and 43 years old (average of 24 years) and they all had self-reported normal hearing. They provided their informed consent before starting the experimental session and were paid for their contribution.

The sample size of 24 participants was assessed a priori. The experiment uses a repeated measures (within-subject) design. It is of interest to check the main effects of two factors: "musical instrument" and "room". The experiment considers 64 sound stimuli that can be grouped into either 8 groups of 8 instrument measurements or 8 groups of 8 room measurements. The first case is of more interest for us, with a null hypothesis that revereberance estimates are the same for the 8 instrument measurements. Based on the simulations shown earlier in this chapter we expect to reject this hypothesis. Assuming a medium effect size (Cohen's $f = 0.25$), an α level (p-value) of 0.05 to support/reject the hypothesis and a power of 90%, 24 participants are required to reach the desired effect size (actual power of 0.96). This analysis was done in the software G*Power (Faul et al., 2007, 2009).

6.6.4 Experimental sessions

The experimental sessions were organised in a one-hour session per participant, including breaks. A multi-stimulus comparison method was used, where the participant was presented with 8 stimuli that he or she had to sort along a scale from 0 to 1 according to an increasing sensation of reverberance. Sixteen trials (i.e., 16 scales with 8 stimuli each) were presented to each participant, with 8 trials having stimuli of the same instrument in different rooms (within-instrument), and 8 trials having different instruments in the same room (within-room).

6.7 Experimental results

6.7.1 Within-instrument evaluation

The experimental results for the within-instrument evaluations are shown in Figure 6.7. The median reverberance estimates $P_{\mathrm{REV,exp}}$ vary between

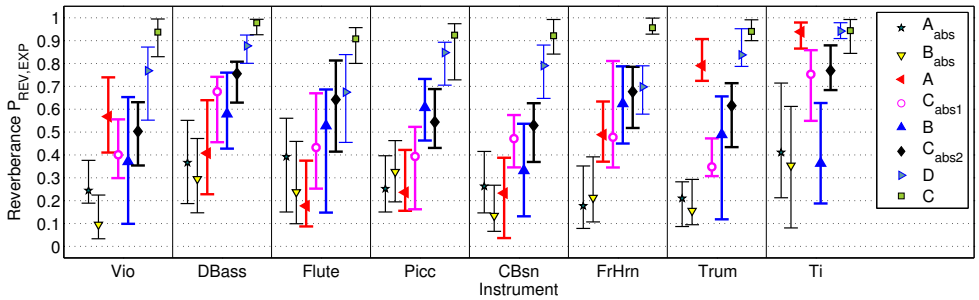


Figure 6.7: Experimental results from the listening test (within-instrument evaluation). The median values of the reverberance estimates in 8 different acoustic environments are indicated together with the interquartile ranges obtained from 24 data points. The eight instruments from left to right are: Vio, DBass, Flute, Picc, CBsn, FrHrn, Trum, and Ti.

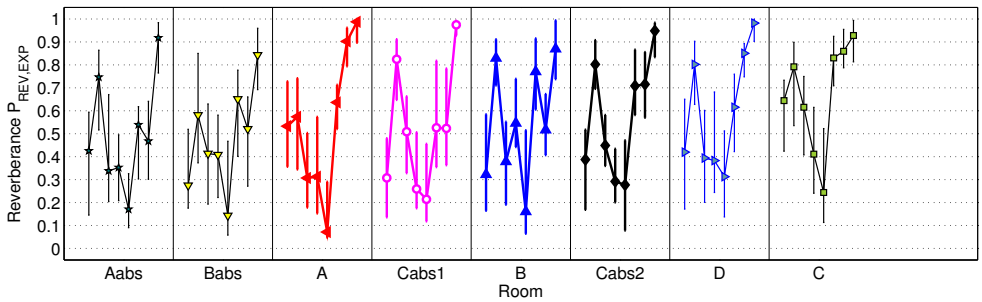


Figure 6.8: Experimental results from the listening test (within-room evaluation). Median values of the reverberance estimates are indicated together with the interquartile ranges obtained from 24 data points. The estimates within each room condition correspond from left to right to: Vio, DBass, Flute, Picc, CBsn, FrHrn, Trum, and Ti.

0.10 (Vio in room B_{abs}) and 0.98 (DBass in room C). Since the sounds were compared within instruments, the individual scales may not be directly related to each other. This is because the participants' responses only required to be referenced to the sound samples within each trial. In the subsequent section, these experimental results per instrument are compared with their corresponding binaural model estimates.

6.7.2 Within-room evaluation

The experimental results for the within-room evaluations are shown in Figure 6.8. The median reverberance estimates vary between 0.07 (CBsn in room A) and 0.99 (Ti in room A). Using the average of the estimated values as an indication of how reverberant the instruments are, the instruments in order of increasing reverberance estimates are: CBsn ($P_{REV,exp}$= 0.20), Picc, Vio, Flute ($P_{REV,exp}$= 0.37 ≈ 0.41 ≈ 0.43), FrHrn, Trum ($P_{REV,exp}$= 0.66 ≈ 0.67), DBass ($P_{REV,exp}$= 0.74), and Ti ($P_{REV,exp}$= 0.93).

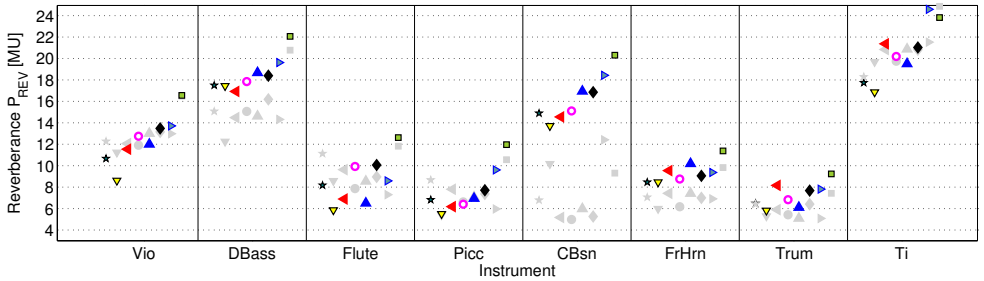


Figure 6.9: New simulated $P_{\mathrm{REV,10\ s}}$ estimates expressed in MU for the eight selected musical instruments in the eight acoustic environments (rooms A-D). For ease of comparison, the corresponding $P_{\mathrm{REV,90\ s}}$ estimates taken from Figure 6.4 are indicated by grey markers.

Table 6.7: Pearson correlation r_p between experimental and simulated P_{REV} estimates in the within-instrument condition. Each r_p value is obtained by comparing 8 pairs of data points (6 degrees of freedom).

	$P_{\mathrm{REV,exp}}$ correlated with		
Nr./ Instrument	$P_{\mathrm{REV,10\ s}}$	$P_{\mathrm{REV,90\ s}}$	$P_{\mathrm{REV,max,90\ s}}$
1/ Vio	0.92*	0.81*	0.77*
7/ DBass	0.85*	0.72*	0.91*
8/ Flute	0.80*	0.22	0.46
9/ Picc	0.90*	0.27	0.26
16/ CBsn	0.93*	0.42	0.77*
17/ FrHrn	0.85*	0.73*	0.90*
20/ Trum	0.90*	0.35	0.74*
22/ Ti	0.89*	0.62**	0.74*

(*) Significant correlation, $p < 0.05$. (**) Correlations that approach significance, $p < 0.10$.

6.8 Comparison between experimental and simulated reverberance estimates

6.8.1 Reference data: New simulations of P_{REV}

The presentation level of the new 10-s excerpts of the orchestra instruments is below the assumed level of the simulations presented in the first part of the chapter, as indicated in the last column ("Δ Pres. Level") of Table 6.6. For this reason, we decided to obtain new P_{REV} estimates using the same instrument excerpts as used in the experimental sessions. The results are shown in Figure 6.9. In the remaining of this chapter, the newly obtained estimates are labelled as $P_{\mathrm{REV,10\ s}}$. In the figure, the reverberance estimates for the 90-s sounds, labelled as $P_{\mathrm{REV,90\ s}}$, are indicated by the grey markers.

Table 6.8: Pearson correlation r_p between the experimental estimates $P_{\mathrm{REV,exp}}$ for all possible instrument pairs. The matrix is symmetric along its diagonal. This table contains mostly high correlation values in contrast to the r_p values of Table 6.5 that have a wider range and even include negative values.

	Instrument							
	Vio	DBass	Flute	Picc	CBsn	FrHrn	Trum	Ti
Vio	-	0.86	0.74	0.78	0.89	0.89	0.95	0.86
DBass	0.86	-	0.92	0.90	0.97	0.91	0.73	0.67
Flute	0.74	0.92	-	0.92	0.92	0.84	0.59	0.41
Picc	0.78	0.90	0.92	-	0.91	0.87	0.68	0.45
CBsn	0.89	0.97	0.92	0.91	-	0.86	0.74	0.69
FrHrn	0.89	0.91	0.84	0.87	0.86	-	0.87	0.67
Trum	0.95	0.73	0.59	0.68	0.74	0.87	-	0.85
Ti	0.86	0.67	0.41	0.45	0.69	0.67	0.85	-

6.8.2 Within-instrument evaluation

The experimental reverberance estimates $P_{\mathrm{REV,exp}}$ of Figure 6.7 can either be compared with (1) P_{REV} estimates computed from the exact 10-s excerpts ($P_{\mathrm{REV,10\,s}}$) used in the experiments, or with (2) the simulated estimates P_{REV} of Figure 6.4, which were obtained for the 90-s excerpts and grouping the same instruments together. The correlation values are shown in Table 6.7. The experimental results are significantly correlated with the $P_{\mathrm{REV,10\,s}}$ values with $r_p(6)$ between 0.80 (Flute) and 0.92 (Vio). When comparing $P_{\mathrm{REV,exp}}$ with $P_{\mathrm{REV,90\,s}}$, only three correlation values (r_p for Vio, DBass, FrHrn) are significant and one approaches significance (r_p for Ti). Although the r_p values are expected to be lower because the $P_{\mathrm{REV,90\,s}}$ estimation considered parts of the sounds that were not presented to the listeners, these estimates could be interpreted as belonging to a more representative playing context of the instruments. Since the selected instruments played *fortissimo* during the 10-s excerpts (taken from bars 10-16 of Brahms Symphony Nr. 4, 3$^{\mathrm{rd}}$ movement) the correlation with the maximum reverberant estimates $P_{\mathrm{REV,max,90\,s}}$ (obtained from the percentile 75 of $P_{\mathrm{REV,90\,s}}$) is also included. In this case, six (of eight) r_p values are significant with values between 0.74 (Trum and Ti) and 0.91 (DBass).

Reverberance trends

The reverberance trends that have been observed in $P_{\mathrm{REV,90\,s}}$ and that may also be observed in the $P_{\mathrm{REV,exp}}$ estimates of each instrument are evaluated by first generating a similarity matrix based on a matrix of correlation values and then using the MDS method to generate a two-dimensional representation. The resulting matrix and Cartesian repre-

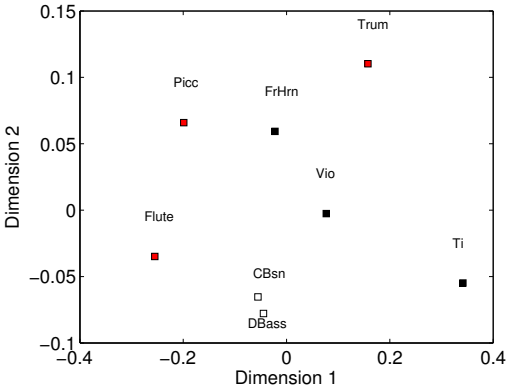


Figure 6.10: Cartesian representation (stress $S_t = 27.9\%$) of the 8 instruments of the orchestra used in the listening experiment. This analysis is based on the correlation matrix shown in Table 6.8. Instruments that are close to each other provide a similar ranking in their P_{REV} values for the eight tested acoustic environments. The instruments are indicated using labels according to the trends found for the simulated reverberance estimates: trend (1), trend (2), and trend (3), which are indicated by the black, red, and white square markers, respectively.

sentation are shown in Table 6.8 and Figure 6.10, respectively. This analysis is not conclusive but it shows that the three instruments from trend (1) (Vio, FrHrn, and Ti) are still mapped in the neighbourhood of each other. In trend (2), Flute and Picc stay near each other but Trum gets farther apart and gets somewhat closer to the French horn (FrHrn) from trend (1). Supported by the simulated $P_{\mathrm{REV,10\ s}}$ values of Figure 6.9, something that all three instruments of trend (2) and FrHrn have in common is their low reverberance estimates. In trend (3), the contrabassoon changed considerably its position with respect to the position shown in Figure 6.5. Its reverberance pattern turned similar to that of the double bass.

We state that the current analysis is "not conclusive" because it is based on the graphical representation shown in Figure 6.10, which has the following limitations:

1) The space has been obtained with a lower number of stimuli (8 instead of 14 as in Figure 6.5). This implicitly assumes that none of the 6 omitted instrument sounds would significantly affect the position of the 8 points that have been obtained.

2) The correlation values r_p in the similarity matrix of Table 6.8 are higher than those of Table 6.5. Only three instruments (Flute, Picc, Ti) have P_{REV} values with at least one r_p value lower than 0.50 with

Table 6.9: Results of repeated measures one-way ANOVAs conducted for each acoustic environment. In all the analyses it was found that the variable "instrument" influences significantly the experimental $P_{\mathrm{REV,exp}}$ values obtained in the within-room evaluations. For each of the eight acoustic environments, 192 observations were available (8 instruments evaluated once by 24 participants).

Room	$F(7,184)$	p	Room	$F(7,184)$	p
A	34.82	< 0.001	C	16.91	< 0.001
$\mathrm{A_{abs}}$	15.52	< 0.001	$\mathrm{C_{abs1}}$	25.96	< 0.001
B	19.61	< 0.001	$\mathrm{C_{abs2}}$	27.94	< 0.001
$\mathrm{B_{abs}}$	10.55	< 0.001	D	22.93	< 0.001

respect to the P_{REV} values of other instruments. This is in contrast with the r_p values of Table 6.5 where all 14 instrument groups have at least one r_p value less than 0.50. This may be an indirect consequence of the reduced number of stimuli (from 8 instruments) in the current analysis.

6.8.3 Within-room evaluation

The within-room results shown in Figure 6.8 can be directly used to evaluate the dependency of reverberance on the sound source type. For room C, which is the most reverberant of the acoustic environments, the instruments sorted from low to high scores, i.e., from least to most reverberant are: CBsn, Picc, Flute, Vio, DBass, FrHrn, Trum, and Ti, respectively. This "reverberance pattern" is similar in the other seven acoustic environments, with a rank-order (Spearman) correlation that ranges between 0.69 (r_s with room B) and 0.98 (r_s with room A).

The average $P_{\mathrm{REV,exp}}$ estimates between 0.20 (for CBsn) and 0.93 (for Ti) may be used as evidence for the dependency of reverberance on the sound source (instrument) type. To provide further statistical evidence, a repeated measures one-way ANOVA (one for each acoustic environment) was conducted to analyse the influence of the variable instrument on $P_{\mathrm{REV,exp}}$. The results show that "instrument type" influenced significantly the reverberance scores in all rooms, as shown in Table 6.9.

6.9 Conclusions

In this chapter we have presented a new implementation of the RAA model which was used to analyse individual instruments of an anechoic orchestra. Those instruments (in total 23 instrument sections, duration of 90 s) were auralised in eight different acoustic conditions having representative reverberation times as found in music performance venues and

rehearsal rooms (0.8-2.5 s). We provide experimental evidence for the validity of the reverberance estimates P_{REV} of the RAA model especially for the case when the same instrument sound is compared in the different acoustic environments.

The reverberance estimates (P_{REV}) obtained from the RAA model varied depending on the spectral content of the analysed instrument and the presentation level. At the same time, we found a large variation of the estimates when using a running analysis window within each individual instrument. The simulated P_{REV} values had a systematic relationship with EDT and T_{30} that could be classified in two different trends explaining 91.3% of the simulated data. In 60.9% of the data, P_{REV} had a higher correlation with EDT than with T_{30}. This trend was found in instruments with a balanced spectrum across critical bands. We could not provide conclusive evidence for the existence of those trends based on the experimental results with 8 (of the 23) instruments. However, the experimental results provided evidence for (1) the significant influence of the instrument type on the perceived reverberation, and (2) the validity of the simulated P_{REV} estimates using sound excerpts that had a duration of 10 s. The simulated P_{REV} estimates were all significantly correlated with the experimental estimates.

Further work is needed to quantify the extent to which reverberance actually depends on the presentation level of the test sounds. In our experimental approach, the presentation level of the instrument sounds was not varied, accounting only for natural level differences due to different sound strength values in each of the acoustic environments. The investigation of this aspect will require further collection of experimental data.

The research presented in this chapter resulted from an exchange (secondment) project at the acoustic consultancy company Müller-BBM. The research goals were: (1) to introduce perception-based predictions of room acoustic indicators to real-world (room) acoustic conditions, and (2) to evaluate to what extent such an approach correlates with listeners' experiences. The significant correlation between simulated and experimental estimates of reverberance is therefore an encouraging result indicating that perception-based predictors are not only of academic interest, but might also improve the predictions obtained in the context of room acoustic consultancy. However, one needs to be aware that such psychoacoustic-based approaches (see also Lee et al., 2012, 2017) represent a fundamental change of paradigm in room acoustics. According

to established measurement guidelines (ISO, 2009), the acoustic properties of a room are considered as (level) linear and time invariant, that is, room properties are assumed to be independent of the type of excitation, and of the level of the exciting signals. Such a source-filter characterisation of room acoustic transmission allows to characterise rooms as linear time-invariant (LTI) systems. The results of this chapter may be used as evidence that the perception of room acoustic parameters (of reverberance, in our case) depends on the context for which the room is used and this is contrary to the idea of an LTI system.

7 General discussion

The work presented in this thesis is concerned with the use of an auditory model for the evaluation of complex sounds, particularly musical instruments, with a special emphasis on the evaluation of perceptual similarity of individual notes played on different pianos. The following instruments have been evaluated: (1) the **hummer** (Chapter 2), which is a simple instrument with sounds that oscillate in amplitude and frequency. Existing recordings and synthesised sounds obtained from a physical model were compared; (2) Recordings of one note played on different **pianos** (Chapter 3, 4, 5), and; (3) Existing recordings of an anechoic **orchestra**, to which the effect of reverberation has been added by digital convolution, generating eight acoustic environments (Chapter 6).

In Chapter 2, sounds of the hummer in its acoustic modes 2 and 4 were evaluated using a selection of psychoacoustic descriptors namely loudness, loudness fluctuations, roughness, and fluctuation strength. An analysis based on fundamental frequency estimates was also included. The analyses were based on reported just-noticeable differences (JND) for each of the 5 evaluated descriptors. The results showed that the synthesised sounds of the hummer are more similar to the recorded ones in acoustic mode 2 than in mode 4. In mode 2, two descriptors had a difference of less than one JND and one descriptor was just above the JND. In mode 4 only one descriptor had a difference of less than one JND. An analysis based on 5 descriptors can be interpreted as an analysis based on 5 "dimensions" that are assumed to be appropriate to evaluate the characteristics of the test sounds.

In Chapter 3 the perceptual similarity among recordings of one note played on different historical Viennese pianos was evaluated. Using the concept of JND, two sounds are perceptually similar along one explicit

"dimension" if they differ by less than one JND. In this chapter perceptual similarity was approached more abstractly, by asking participants to discriminate two sounds while modifying the degree of similarity between them. The objective was to develop a method where the similarity between sounds can not only be assessed but also manipulated by using a specifically generated noise. The noise used to manipulate the difficulty of the task follows the spectro-temporal properties of the sounds being tested and is derived from a modified ICRA noise algorithm. The experimental method, that we named instrument-in-noise, was compared with the method of triadic comparisons, which is a widely used method to evaluate the similarity among stimuli. For similarity estimates using 7 piano sounds, the correlation between the results of both methods was $r_p(17) = -0.47$, $p = 0.04$, and $r_s(19) = -0.64$, $p < 0.001$. We concluded that the instrument-in-noise method is a promising method to evaluate the similarity between sounds.

In Chapter 4 the instrument-in-noise method is simulated using an existing computational model of auditory processing. The auditory (PEMO) model developed by Dau et al. (1997a) was used. The model was described together with the choice of parameters for each of its stages. The model uses a back-end decision stage (central processor) that processes the outputs of the model, i.e., the internal representations of the incoming sounds. We developed a custom implementation of the central processor to enable the artificial listener (i.e., the model) to estimate the amount of noise needed to correctly discriminate two piano sounds. We used the same piano sounds and ICRA noises as in Chapter 3. The simulated and experimental thresholds had a moderate to high correlation with $r_p(17) = 0.54$, $p = 0.02$, and $r_s(19) = 0.63$, $p < 0.001$.

In Chapter 5 the instrument-in-noise method was further evaluated using the same set of piano sounds to which the reverberation of a large room (ground area of 570 m^2 and EDT of 3.0 s at mid frequencies) was added by means of digital convolution. The instrument-in-noise method was evaluated experimentally (similar to Chapter 3) and by running simulations (similar to Chapter 4). The results of this chapter showed that: (1) For the **experimental data**, thresholds of the instrument-in-noise method $\text{thres}_{\text{exp}}$ are correlated with the results of the experimental triadic comparisons with $r_p(18) = -0.49$, $p = 0.03$, and $r_s(19) = -0.65$, $p < 0.001$; (2) For the obtained **instrument-in-noise thresholds**, the experimental $\text{thres}_{\text{exp}}$ and simulated $\text{thres}_{\text{sim}}$ values are correlated with $r_p(18) = 0.58$, $p < 0.01$, and $r_s(19) = 0.61$, $p < 0.001$.

In Chapter 6, an example of the auditory modelling framework applied to room acoustics is given. More specifically, a binaural auditory (RAA) model (van Dorp, 2011) is used to study the perceived reverberation (reverberance) of different instrument sounds in eight simulated rooms. The RAA model has peripheral stages similar to the PEMO model that are applied independently to left- and right-ear signals, and the central processor converts individual internal representations into a metric of reverberance P_{REV}. Listening experiments with 8 of the instruments were conducted to test the validity of the RAA model in a within-instrument modality (same instrument evaluated in the eight rooms) and in a within-room evaluation (same room for eight different instruments). The results of the within-instrument evaluation showed that P_{REV} estimates are highly correlated with experimental estimates having $r_p(6)$ values ranging between 0.80 and 0.93. The experimental results of the within-room evaluation showed that in all the environments the instrument type (i.e., sound source type) influences significantly the participants' reverberance scores. The extension of the use of the unified modelling framework of Chapters 4 and 5 to this application by just adopting a different but "suitable" central processor stage shows the potential of using psychoacoustic modelling in auditory tasks that are different to those for which the models have been previously validated (see, e.g., Appendix D).

7.1 Advantages of the current auditory modelling approach

Experience was gained on the perceptual modelling of a listening task, namely the instrument-in-noise method, that was designed to evaluate the similarity among sounds (Chapter 3). Our implementation of the task can provide interesting information about the sounds being evaluated. Some of these benefits are listed in this section.

The instrument-in-noise method was implemented to compare pairs of sounds using a 3-AFC task. An auditory model was used to produce internal representations of the three sequentially-presented test intervals upon which the artificial listener chose one as being (most likely) different to the other two test intervals. One of the primary advantages of this approach is, therefore, the possibility to **algorithmically evaluate perceptual aspects** of the sounds being compared.

One example of algorithmic evaluation was presented in Chapter 6.

In that chapter an existing auditory model was used to simulate the perceived reverberation (reverberance) elicited by a set of different instrument sounds. In a listening experiment presented in the second part of the chapter we assumed that instrument sounds for which the "artificial listener" provided similar reverberance estimates were also going to be judged as similar by human listeners. Motivated by this idea, we chose a subset of 8 (of 23) instrument sounds for which the auditory model showed a characteristic reverberance performance (different trends). Hence, the model simulations (first part of the chapter) were used as a way to obtain some "a priori" knowledge about human performance, helping with the design of the listening experiment.

Another interesting aspect of the **internal representations** obtained from the auditory model is that they are **multidimensional**. The dimensions of the representations are related to time, audio frequency, and modulation frequency. Therefore the current approach provides the possibility to perform an advanced "sound feature analysis" based on information available along either of those dimensions. Since the objective of this thesis was to use the auditory modelling framework in a similarity task, our "advanced analysis" of the multidimensional piano representations was used to investigate which cues along the three dimensions may have been used by the artificial listener (and potentially also by our participants) to judge the piano sounds, rather than looking at what physical properties of the piano sounds lead to such representations. A complementary approach where piano sounds have been analysed in terms of sound features is given by Chaigne, Osses, and Kohlrausch (2018). In that study, four of the C#$_5$-piano sounds used in Chapter 3 (P4-P7) were evaluated together with recordings of other five notes (C$_2$, F$_3$, C$_4$, A$_4$, G$_6$). A comparison between the results of our information-based analysis and their seven spectro-temporal descriptors may provide further insights into how the physical properties of the piano are actually related to perceptual aspects.

7.2 Limitations of the current approach

The auditory (PEMO) model has been applied to the specific case of similarity between sounds (Chapter 4 and 5). We identified a number of limitations of our approach that are related to (1) the choice of the auditory model, (2) the way the similarity task was implemented, and (3, 4) the way the information of the optimal detector was limited and reduced.

7.2.1 Choice of the model

The PEMO model used in this thesis has a **level-independent critical-band filter bank** (stage 2, ERB filter bank in Figure 4.1, page 56). This is in contrast to the non-linear behaviour (compressive characteristic) of the basilar membrane, which is more compressive towards higher frequencies (see, e.g., Saremi et al., 2016, their Figure 3). Given that we found the decision criterion of the piano discrimination with notes of the same pitch to rely mostly on a frequency region above F0, particularly between 1000 and 3000 Hz, comprising about 4 harmonics of the note (see panel (a) of Figure 4.10, page 77), the use of a non-linear critical-band filter bank coupled to the auditory model would change the sensitivity of the model to our piano samples, that would in turn affect the estimation of simulated thresholds. Our motivation to choose the PEMO model and, therefore, the Gammatone filter bank came from our higher degree of success in replicating simulated data reported in the literature compared with more recent versions of the auditory model.

The PEMO model was used as a **monaural model** despite the fact that the piano sounds were presented diotically (Chapter 3) and binaurally (Chapter 5) to the participants. We would not expect significant changes between monaural and diotic discrimination thresholds (see, e.g., Langhans & Kohlrausch, 1992) and although we did not use the right-ear channels of the piano sounds in the simulations of Chapter 5, we would expect that similar discrimination cues are available with respect to the use of the left-ear channel. In order to further apply the PEMO model to other auditory tasks it is important to evaluate the role of processing left and right-ear signals in parallel and by coupling their internal representations to have access to binaural cues as Breebaart et al. (2001) did. The use of the PEMO model in such a context would allow the use of modulation-frequency information for simulating binaural tasks.

7.2.2 Implementation of the similarity task

The similarity task was implemented as a 3-AFC discrimination experiment. With this approach, the test sounds are **presented sequentially to the participants** and the similarity assessment is based on the comparison of **individual piano notes** that have the same F0 and the same duration. Due to the implementation of the task as sequentially-presented intervals, a simple top-down approach (memory templates) could be adopted, assuming that the participant is able to "learn" and use this information always in the best possible way. In practice, this

represents a situation where the participant is recursively exposed and gets familiar with the sounds. Hence, the presentation of sounds as individual notes represents a condition where the participant can focus on smaller sound differences compared to, e.g., melodic lines with multiple notes (and/or multiple instruments) where there is less exposure to one individual note (and/or instrument). In that case a more elaborate top-down approach would be needed. Such an approach should use some sort of information weighting that may be influenced by attention and/or saliency aspects.

7.2.3 Additive internal noise

The internal noise was used to limit the artificial listener's performance in an intensity-discrimination task (see Appendix D). The use of such a simple additive internal noise was found to be not accurate enough in simulations of several AM detection tasks (Ewert & Dau, 2004). To overcome this limitation, Wallaert, Moore, Ewert, and Lorenzi (2017) adopted a multiplied noise as an additional source of internal variability besides the additive internal noise and a memory noise they used to reduce the memory capacity of the model (that can be compared with our use of t_{obs}).

7.2.4 Reduction of information in the optimal detector

The artificial listener was found to be too sensitive to differences in the stimuli when considering whole-duration piano waveforms with t_{obs} durations of 1.5 and 2.2 s in Chapters 4 and 5, respectively. As a way to **reduce available cues** in the model, shorter observation durations t_{obs} were evaluated with as result t_{obs} values of 0.25 and 0.20 s. We did not evaluate other forms of information reduction such as the application of (additional) smoothing to the internal representations or the use of a temporal weighting that could provide a higher emphasis to the information present in the first 0.20-0.25 s with respect to the rest of the representation instead of removing the latter one completely.

7.3 Perspectives for further research

The modelling framework used in this thesis includes stages of peripheral processing of the auditory system and provides the possibility to add a back-end stage or central processor. To apply such an approach to a similarity task involving piano sounds we had to (1) choose the appropriate parameters to be used in the peripheral processing part, and to

(2) adjust the central processor in a way that two or more sound (internal) representations could be conveniently compared to each other to assess how similar they are. This corresponds to a very general approach and we believe that it can be applied to many other applications as long as "hearing" is involved. We give next two examples of potential applications, one related to room acoustics and another related to human echolocation. For both examples it would be desirable to use the auditory model in a **binaural set-up**, using a suitable coupling of left and right-ear internal representations in the central processor.

The first example of application was actually given in Chapter 6 where a binaural model (the RAA model) was used to investigate the reverberance of different sound sources in room acoustics. The particular context in which that chapter was developed was a consequence of a joint project with the acoustic consultancy company Müller-BBM. The goal of the project was to use an existing binaural model in the evaluation of recorded (auralised) sounds in different rooms, evaluating to what extent reverberance estimates from the model correlate with physical measurements of reverberation time using standardised procedures (ISO, 2009). Our goal was, therefore, to evaluate how well did the (existing) RAA model perform rather than pursuing an improvement of the simulation power of that model. The use of this psychoacoustic-based model suggests a change in paradigm in room acoustics. The ISO procedures encourage the characterisation of an acoustic environment independent of the sound source and its level, which is in contrast to the approach of using the RAA model along with, e.g., the use of loudness-based reverberation estimates (Lee et al., 2012). In this context, we suggest two possible ways to further extend the use of the binaural RAA model: (1) To investigate the dependency of reverberance on the presentation level of the stimuli. This is motivated by the **strong level dependency** that we identified in the model –also recently reported by Lee et al. (2017)– and requires further experimental evidence; and (2) to extend the validation of the RAA model to other room acoustics parameters, such as clarity, listener envelopment, and apparent source width (van Dorp, 2011; van Dorp et al., 2013) using more sound sources.

Our second example of potential application is the use of (binaural) auditory modelling to study human echolocation. Echolocation is a perceptual ability mostly used by blind people to explore a given spatial environment. Sounds that are emitted orally ("source") are normally used by them to extract information about surrounding objects (in a

"medium") based on the spectral and spatial cues conveyed in the sounds that are heard back ("receiver"). Experiments on human echolocation have been mostly implemented as performance tasks (see, e.g., Dufour et al., 2005; Guzmán, 2016; Rowan et al., 2017) which, based on the arguments presented in the introduction of this thesis, is an auspicious condition to be simulated by means of auditory models. Two types of echolocation tasks are the localisation of an object and the discrimination of the size of an object (de Vos & Hornikx, 2018). Data from such tasks analysed using an information-based approach of the underlying internal representations as used in Chapters 4 and 5 may provide useful insights to optimise the "sound source" by, e.g., developing artificially generated (optimal) clicks, or to optimise the "medium" by designing rooms that enhance the transmission of spectral and spatial cues.

7.4 General conclusion

The main goal of this thesis was to gain insights into the perceptual modelling of "an" auditory task. We focused our efforts on the perceptual similarity of a specific note ($C\#_5$) played on a set of 7 historical Viennese pianos by using an auditory model. For doing this we developed a method where the similarity between two sounds could be manipulated by using noise, allowing to evaluate similarity as a performance task. The method, that we named instrument-in-noise, was compared with the method of triadic comparisons reaching moderate to high correlations using the piano sounds in two acoustic conditions: "anechoic" and reverberant (EDT of 3 s). An existing modelling approach based on a model of the effective processing in the auditory system was used. The simulated thresholds $\text{thres}_{\text{sim}}$ were in both cases highly correlated with the experimental thresholds $\text{thres}_{\text{exp}}$, but they had a strong "primacy" effect, where only the first 0.25 or 0.20 s of the internal representations were used to produce these results. The encouraging results of our modelling approach allowed us to perform information-based analyses on the piano internal representations. We concluded that the weighting of information used by the artificial listener may be similar to that used by human listeners. The advantages and limitations of both experimental and modelling approach were discussed. Due to the use of the unified auditory modelling framework offered by the adopted model, further research is suggested in applications involving binaural listening, which represents a different type of auditory task to that implemented here for the perceptual similarity between stimuli.

References

Agus, T., Suied, C., Thorpe, S., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *J. Acoust. Soc. Am.*, *131*(5), 4124–4133.

Aures, W. (1985). Ein Berechnungsverfahren der Rauhigkeit. *Acustica*, *58*(5), 268–281.

Beranek, L. (2004). *Concert halls and opera houses: Music, acoustics and architecture*. Springer New York.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, *17*, 97–110.

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot Int.*, *5*(9/10), 341–345.

Bradley, R. (1953). Some statistical methods in taste testing and quality evaluation. *Biometrics*, *9*(1), 22–38.

Breebaart, J., van de Par, S., & Kohlrausch, A. (2001). Binaural processing model based on contralateral inhibition. I. Model structure. *J. Acoust. Soc. Am.*, *110*(2), 1074–1088.

Burton, M., & Nerlove, S. (1976). Balanced designs for triads tests: Two examples from English. *Social Science Research*, *5*, 247–267.

Carroll, J., & Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika*, *35*(3), 283–319.

Chabassier, J., Chaigne, A., & Joly, P. (2013). Modeling and simulation of a grand piano. *J. Acoust. Soc. Am.*, *134*(1), 648–65.

Chaigne, A. (2016). Acoustics of pianos: An historical perspective. In *International Symposium on Musical and Room Acoustics.* La Plata.

Chaigne, A., Hennet, M., Chabassier, J., & Duruflé, M. (2016). Comparison between three different Viennese pianos of the nineteenth century. In *International Congress on Acoustics.* Buenos Aires.

Chaigne, A., Osses, A., & Kohlrausch, A. (2018). Similarity of piano tones: a psychoacoustical and sound analysis study. *Applied Acoustics (submitted).*

Chalupper, J., & Fastl, H. (2002). Dynamic loudness model (DLM) for normal and hearing-impaired listeners. *Acta Acust. united Ac.*, *88*(3), 378–386.

Chelazzi, L., Miller, E., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, *363*, 345.

Daniel, P., & Weber, R. (1997). Psychoacoustical roughness: Implementation of an optimized model. *Acustica - Acta Acustica*, *83*, 113–123.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J. Acoust. Soc. Am.*, *102*(5), 2892–2905.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *J. Acoust. Soc. Am.*, *102*(5), 2906–2919.

Dau, T., Püschel, D., & Kohlrausch, A. (1996a). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.*, *99*(6), 3615–3622.

Dau, T., Püschel, D., & Kohlrausch, A. (1996b). A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements. *J. Acoust. Soc. Am.*, *99*(6), 3623–3631.

Dau, T., Verhey, J., & Kohlrausch, A. (1999). Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers. *J. Acoust. Soc. Am.*, *106*(5), 2752–2760.

De Man, B., & Reiss, J. (2013). A pairwise and multiple stimuli approach to perceptual evaluation of microphone types. In *Audio Engineering Society Convention 134.* Rome, Italy.

De Man, B., & Reiss, J. (2014). APE: Audio Perceptual Evaluation Toolbox for MATLAB. In *Audio Engineering Society Convention 136.* Berlin, Germany.

de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization. *J. Stati. Softw.*, *31*(3), 1–30.

Derveaux, G., Chaigne, A., Joly, P., & Becache, E. (2003). Time-domain simulation of a guitar: Model and method. *J. Acoust. Soc. Am.*, *114*(6), 3368–3383.

de Vos, R., & Hornikx, M. (2018). Human ability to judge relative size and lateral position of a sound reflecting board using click signals: Influence of source position and click properties. *Acta Acust. united Ac.*, *104*, 131–144.

Dreschler, W., Verschuure, H., Ludvigsen, C., & Westermann, S. (2001). ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. *Int. J. Audiol.*, *40*(3), 148–157.

Drullman, R., Festen, J., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, *95*, 1053.

Dubois, D. (2000). Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive Science Quarterly*, *1*, 35–68.

Dufour, A., Després, O., & Candas, V. (2005). Enhanced sensitivity to echo cues in blind subjects. *Exp. Brain Res.*, *165*, 515–519.

Ellis, D. (2002). *A phase vocoder in MATLAB.* Retrieved from http://www.ee.columbia.edu/{~}dpwe/resources/matlab/pvoc/

Everitt, B. (2005). Multidimensional scaling and correspondence analysis. In *An R and S-PLUS companion to multivariate analysis* (pp. 91–114). Springer Verlag.

Ewert, S. (2013). AFC - A modular framework for running psychoacoustic experiments and computational perception models. In *Proceedings of the International Conference on Acoustics AIA-DAGA* (pp. 1326–29).

Ewert, S., & Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *J. Acoust. Soc. Am.*, *108*(3), 1181–1196.

Ewert, S., & Dau, T. (2004). External and internal limitations in amplitude-modulation processing. *J. Acoust. Soc. Am.*, *116*(1), 478–490.

Fastl, H. (1977). Roughness and temporal masking patterns of sinusoidally amplitude-modulated broadband noise. In E. Evans & J. Wilson (Eds.), *Psychophysics and physiology of hearing* (pp. 403–417). Academic Press.

Fastl, H. (1982). Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise. *Hear. Res.*, *8*(1), 59–69.

Fastl, H. (1983). Fluctuation strength of modulated tones and broadband noise. In *Hearing - physical bases and psychophysics* (pp. 282–288).

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics, Facts and Models* (Third ed.). Springer Berlin Heidelberg.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods*, *39*(2), 175–191.

Fog, C., & Ballinger, R. (2008). A new symphonic hall, Musikhuset Aarhus, Denmark. In *Proceedings of acoustics 08 paris* (pp. 363–368).

Francart, T., van Wieringen, A., & Wouters, J. (2008). APEX 3: a multi-purpose test platform for auditory psychophysical experiments. *J. Neurosci. Meth.*, *172*(2), 283–93.

Fritz, C., Cross, I., Moore, B., & Woodhouse, J. (2007). Perceptual thresholds for detecting modifications applied to the acoustical properties of a violin. *J. Acoust. Soc. Am.*, *122*(6), 3640–3650.

Fritz, C., & Dubois, D. (2015). Perceptual evaluation of musical instruments: state of the art and methodology. *Acta Acust. united Ac.*, *101*, 369–381.

Fritz, C., Woodhouse, J., Cheng, F., Cross, I., Blackwell, A., & Moore, B. (2010). Perceptual studies of violin body damping and vibrato. *J. Acoust. Soc. Am.*, *127*(1), 513–524.

García, R. (2015). *Modelling the sensation of fluctuation strength* (Master thesis). Eindhoven University of Technology.

Genuit, K. (1997). Background and practical examples of sound design. *Acustica - Acta Acustica*, *83*(5), 805–812.

Glasberg, B., & Moore, B. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, *47*, 103–138.

Glasberg, B., & Moore, B. (2002). A model of loudness applicable to time-varying sounds. *J. Audio Eng. Soc.*, *50*(5), 331–342.

Goode, R., Killion, M., Nakamura, K., & Nishihara, S. (1994). New knowledge about the function of the human middle ear: Development of an improved analog model. *Am. J. Otol.*, *15*(2), 145–154.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. John Wiley & Sons Inc.

Greenwood, D. (1990). A cochlear frequency position function for several species–29 years later. *J. Acoust. Soc. Am.*, *87*(6), 2592–2605.

Grey, J. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, *61*(5), 1270–1277.

Grey, J. (1978). Timbre discrimination in musical patterns. *J. Acoust. Soc. Am.*, *64*(2), 467–472.

Grey, J., & Gordon, J. (1978). Perceptual effects of spectral modifications on musical timbres. *J. Acoust. Soc. Am.*, *63*(5), 1493–1500.

Guastavino, C., & Katz, B. (2004). Perceptual evaluation of multidimensional spatial audio reproduction. *J. Acoust. Soc. Am.*, *116*(2), 1105–1115.

Guzmán, R. (2016). *The effects of multiple sound reflections on human echolocation: Acoustical analysis of binaural cues in different rooms* (Master thesis). University of Southampton.

Hansen, M., & Kollmeier, B. (2000). Objective modeling of speech quality with a psychoacoustically validated auditory model. *J. Audio Eng. Soc.*, *14*(6), 395–409.

Hirschberg, M., Rudenko, O., Nakiboğlu, G., Holten, A., Willems, J., & Hirschberg, A. (2013). The voice of the mechanical dragon. In *Proceedings of SMAC.* Stockholm.

Hohmann, V. (2002). Frequency analysis and synthesis using a Gammatone filterbank. *Acust. Acta Acust.*, *88*(3), 433–442.

Holube, I., & Kollmeier, B. (1996). Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *J. Acoust. Soc. Am.*, *100*(3), 1703–1716.

Houtsma, A., Rossing, T., & Wagenaars, W. (1987). *Auditory demonstrations.* Eindhoven: Acoustical Society of America.

Huber, R., & Kollmeier, B. (2006). PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception. *IEEE Trans. Audio, Speech, Lang. Process.*, *14*(6), 1902–1911.

ISO. (2009). *ISO 3382-1:2009. Acoustics. Measurement of room acoustic parameters – Part 1: Performance spaces.*

ITU-R. (2015). *BS.1534-3: Method for the subjective assessment of intermediate quality level of coding systems.*

Jepsen, M., Ewert, S., & Dau, T. (2008). A computational model of human auditory signal processing and perception. *J. Acoust. Soc. Am.*, *124*(1), 422–438.

Jillings, N., De Man, B., Moffat, D., Reiss, J., & Stables, R. (2016). Web Audio Evaluation Tool: A framework for subjective assessment of audio. *2nd Web Audio Conference*.

Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J. Acoust. Soc. Am.*, *130*(3), 1475–1487.

Kates, J., & Arehart, K. (2014). The hearing-aid speech quality index (HASQI) version 2. *J Audio Eng. Soc.*, *62*, 99–117.

Kemp, S. (1982). Roughness of frequency-modulated tones. *Acustica*, *50*, 126–133.

Kingdom, F., & Prins, N. (2016). *Psychophysics: A practical introduction* (2nd ed.). Elsevier.

Klockgether, S., & van de Par, S. (2014). A Model for the Prediction of Room Acoustical Perception based on the Just Noticeable Differences of Spatial Perception. *Acta Acust. united Ac.*, *100*, 964–971.

Klockgether, S., & van de Par, S. (2016). Just noticeable differences of spatial cues in echoic and anechoic acoustical environments. *J. Acoust. Soc. Am.*, *140*(4), EL352–EL357.

Kohlrausch, A., Braasch, J., Kolossa, D., & Blauert, J. (2013). An introduction to binaural processing. In *The technology of binaural listening* (pp. 1–32). Springer Berlin Heidelberg.

Kohlrausch, A., Fassel, R., & Dau, T. (2000). The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. *J. Acoust. Soc. Am.*, *108*(2), 723–734.

Kohlrausch, A., Hermes, D., & Duisters, R. (2005). Modeling roughness perception for sounds with ramped and damped temporal envelopes. *Forum Acusticum*(1), 1719–1724.

Kohlrausch, A., Püschel, D., & Alphei, H. (1992). Temporal resolution and modulation analysis in models of the auditory system. In M. Schouten (Ed.), *The auditory processing of speech* (Vol. 10, pp. 85–98). Mouton de Gruyter.

Kruskal, J. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kruskal, J. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29*(2), 115–129.

Langhans, A., & Kohlrausch, A. (1992). Differences in auditory performance between monaural and dichotic conditions. I: masking thresholds in frozen noise. *J. Acoust. Soc. Am.*, *91*(6), 3456–3470.

Langner, G., & Schreiner, C. (1988). Periodicity coding in the Inferior Colliculus of the cat. I. Neuronal mechanisms. *J. Neurophysiol.*, *60*(6), 1799–1822.

Lee, D., Cabrera, D., & Martens, W. (2012). The effect of loudness on the reverberance of music: Reverberance prediction using loudness models. *J. Acoust. Soc. Am.*, *131*(2), 1194–1205.

Lee, D., van Dorp, J., Cabrera, D., & Qiu, X. (2017). Comparison of psychoacoustic-based reverberance parameters. *J. Acoust. Soc. Am.*, *142*(4), 1832–1840.

Leong, V., Stone, M. a., Turner, R., & Goswami, U. (2014). A role for amplitude modulation phase relationships in speech rhythm perception. *J. Acoust. Soc. Am.*, *136*(1), 366–381.

Levelt, W., van de Geer, J., & Plomp, R. (1966). Triadic comparisons of musical intervals. *Br. J. Math. Stat. Psychol.*, *19*, 163–179.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, *49*(2), 467–477.

Lopez-Poveda, E., & Meddis, R. (2001). A human nonlinear cochlear filterbank. *J. Acoust. Soc. Am.*, *110*(6), 3107–3118.

Mao, J., & Carney, L. (2015). Tone-in-noise detection using envelope cues: Comparison of signal-processing-based and physiological models. *J. Assoc. Res. Otolaryngol.*, *16*(1), 121–133.

McAdams, S., & Bigand, E. (Eds.). (1993). *Thinking in sound: The cognitive psychology of human audition*. Oxford University Press.

Meddis, R., & Hewitt, M. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J. Acoust. Soc. Am.*, *89*(6), 2866–2882.

Meddis, R., & O'Mard, L. (1997). A unitary model of pitch perception. *J. Acoust. Soc. Am.*, *102*(3), 1811–1820.

Meyer, J. (2009). *Acoustics and the performance of music*. Springer.

Miller, G. (1947). Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *J. Acoust. Soc. Am.*, *19*, 609.

Moore, B. (2003). Temporal integration and context effects in hearing. *J. Phonetics*, *31*, 563–574.

Moore, B., & Glasberg, B. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.*, *74*(3), 750–753.

Moore, B., & Sek, A. (1992). Detection of combined frequency and amplitude modulation. *J. Acoust. Soc. Am.*, *92*(6), 3119–3131.

Münkner, S. (1993). *Modellentwicklung und Messungen zur Wahrnehmung nichtstationärer akustischer Signale* (Ph.D. thesis). University of Göttingen.

Nakiboğlu, G., Rudenko, O., & Hirschberg, A. (2012). Aeroacoustics of the swinging corrugated tube: Voice of the Dragon. *J. Acoust. Soc. Am.*, *131*(1), 749–765.

Novello, A., McKinney, M., & Kohlrausch, A. (2011). Perceptual evaluation of inter-song similarity in Western popular music. *Journal of New Music Research*, *40*(1), 1–26.

Osses, A., García, R., & Kohlrausch, A. (2016). Modelling the sensation of fluctuation strength. *Proc. Mtgs. Acoust.*, *28*(050005), 1–8.

Patel, A., Iversen, J., & Rosenberg, J. (2006). Comparing the rhythm and melody of speech and music: the case of British English and French. *J. Acoust. Soc. Am.*, *119*(5), 3034–3047.

Patterson, R. (1976). Auditory filter shapes derived with noise stimuli. *J. Acoust. Soc. Am.*, *59*(3), 640–654.

Pralong, D., & Carlile, S. (1996). The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *J. Acoust. Soc. Am.*, *100*(6), 3785–3793.

Püschel, D. (1988). *Prinzipien der zeitlichen der Analyse beim Hören* (Ph.D. thesis). University of Göttingen.

Raake, A., Wierstorf, H., & Blauert, J. (2014). A case for TWO!EARS in audio quality assessment. In *Forum Acusticum* (pp. 1–10). Krakow.

Rabinowitz, W. (1970). *Frequency and intensity resolution in audition* (Master thesis). Massachusetts Institute of Technology.

Rindel, J. (2015). *Orchestra simulation and auralisation.* Odeon. Retrieved from http://www.odeon.dk/pdf/Application{_}note{_}Orchestra{_}auralisation.pdf

Robles, L., & Ruggero, M. (2001). Mechanics of the mammalian cochlea. *Physiol. Rev.*, *81*(3), 1305–1352.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Phil.Trans. R. Soc. London*, *336*, 367–373.

Rowan, D., Papadopoulos, T., Archer, L., Goodhew, A., Cozens, H., Guzmán, R., Edwards, D., Holmes, H., & Allen, R. (2017). The detection of virtual objects using echoes by humans: Spectral cues. *Hear. Res.*, *350*, 205–216.

Saitis, C., Fritz, C., Guastavino, C., & Scavone, G. (2013). Conceptualization of violin quality by experienced performers. In *Proceedings of SMAC* (pp. 123–128). Stockholm.

Saitis, C., Fritz, C., Scavone, G., Guastavino, C., & Dubois, D. (2017). Perceptual evaluation of violins: A psycholinguistic analysis of preference verbal descriptions by experienced musicians. *J. Acoust. Soc. Am.*, *141*(4), 2746–2757.

Saldanha, E., & Corso, J. (1964). Timbre cues and the identification of musical instruments. *J. Acoust. Soc. Am.*, *36*, 2021–2026.

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., & Verhulst, S. (2016). A comparative study of seven human cochlear filter models. *J. Acoust. Soc. Am.*, *140*(3), 1618–1634.

Schlittmeier, S. J., Weissgerber, T., Kerber, S., Fastl, H., & Hellbrück, J. (2012). Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength. *Atten. Percept. Psychophys.*, *74*(1), 194–203.

Schroeder, M. (1968). Reference signal for signal quality studies. *J. Acoust. Soc. Am.*, *44*(6), 1735–1736.

Shannon, R., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Shepard, R. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Søndergaard, P., & Majdak, P. (2013). The Auditory Modeling Toolbox. In J. Blauert (Ed.), *The technology of binaural listening* (pp. 33–56). Springer Berlin Heidelberg.

Sontacchi, A. (1998). *Entwicklung eines Modulkonzeptes für die psychoakustische Geräuschanalyse unter MATLAB* (Unpublished doctoral dissertation). Technischen Universität Graz.

Steeneken, H. (1992). *On measuring and predicting speech intelligibility* (Ph.D. thesis). University of Amsterdam.

Stevens, S. (1955). The measurement of loudness. *J. Acoust. Soc. Am.*, *27*(5), 815–829.

Stevens, S. (1956). The direct estimation of sensory magnitudes–loudness. *Am. J. Psychol.*, *69*(1), 1–25.

Tahvanainen, H., Pätynen, J., Lokki, T., Tahvanainen, H., Pätynen, J., & Lokki, T. (2015). Studies on the perception of bass in four concert halls. *Psychomusicology: Music, Mind, and Brain*, *25*(3), 294–305.

Teret, E., Pastore, T., & Braasch, J. (2017). The influence of signal type on perceived reverberance. *J. Acoust. Soc. Am.*, *141*(3), 1675–1682.

Terhardt, E. (1978). Psychoacoustic evaluation of musical sounds. *Perception & Psychophysics*, *23*(6), 483–92.

Terhardt, E. (1979). Calculating virtual pitch. *Hear. Res.*, *1*, 155–182.

van de Par, S., & Kohlrausch, A. (1995). Analytical expressions for the envelope correlation of narrow-band stimuli. *J. Acoust. Soc. Am.*, *98*(6), 3157–3169.

van Dorp, J. (2011). *Auditory modelling for assessing room acoustics* (Ph.D. thesis). Technische Universiteit Delft.

van Dorp, J., de Vries, D., & Lindau, A. (2013). Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model. *J. Acoust. Soc. Am.*, *133*(3), 1572–1585.

van Veen, T., & Houtgast, T. (1983). On the perception of spectral modulations. In *Hearing - physical bases and psychophysics* (p. 277).

Vogel, A. (1975). Über den Zusammenhang zwischen Rauhigkeit und Modulationsgrad (On the relation between roughness and degree of modulation). *Acustica*, *32*(5), 300–306.

von Klitzing, R., & Kohlrausch, A. (1994). Effect of masker level on overshoot in running- and frozen-noise maskers. *J. Acoust. Soc. Am.*, *95*(4), 2192–2201.

Wallaert, N., Moore, B., Ewert, S., & Lorenzi, C. (2017). Sensorineural hearing loss enhances auditory sensitivity and temporal integration for amplitude modulation. *J. Acoust. Soc. Am.*, *141*(2), 971–980.

Westerman, L., & Smith, R. (1984). Rapid and short-term adaptation in auditory nerve responses. *Hear. Res.*, *15*(3), 249–260.

Wickelmaier, F., & Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments & Computers*, *36*(1), 29–40.

Widmann, U. (1997). Three application examples for sound quality design using psychoacoustic tools. *Acustica - Acta Acustica*, *83*(5), 819–826.

Yang, M., & Kang, J. (2013). Psychoacoustical evaluation of natural and urban sounds in soundscapes. *J. Acoust. Soc. Am.*, *134*(1), 840–51.

Yost, W., Braida, L., Hartmann, W., Kidd, G., Kruskal, J., Pastore, R., Sachs, M., Sorkin, R., & Warren, R. (1989). *Classification of complex nonspeech sounds* (Tech. Rep.). Washington D.C.: National Academy.

Zhou, T., Zhang, M., & Li, C. (2015). A model for calculating psychoacoustical fluctuation strength. *J. Audio Eng. Soc.*, *63*(9), 713–724.

Zilany, M., Bruce, I., Nelson, P., & Carney, L. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *J. Acoust. Soc. Am.*, *126*(5), 2390–2412.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *J. Acoust. Soc. Am.*, *33*(2), 248.

Zwicker, E. (1977). Procedure for calculating loudness of temporally variable sounds. *J. Acoust. Soc. Am.*, *62*(3), 675–682.

Zwicker, E., Flottorp, G., & Stevens, S. (1957). Critical band width in loudness summation. *J. Acoust. Soc. Am.*, *29*(5), 548–557.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, *68*(5), 1523–1525.

List of figures

List of tables

Appendices

The following appendices are included in the next pages:

A. **Auditory frequency scales**
This appendix contains a summary of two auditory frequency scales that are inspired by the concept of critical-bands. The two scales are (1) the critical-band rate z in Bark, which is used in Chapter 2, and (2) the ERB-scale expressed in ERB numbers, which is used in the remaining chapters.

B. **Model of fluctuation strength**
This appendix contains the computational model of fluctuation strength as used in Chapter 2.

C. **Adaptation loops**
This appendix contains an in-depth description of the underlying properties of the adaptation loops used in the auditory models. Both the PEMO (Chapter 4) and RAA models (Chapter 6) include an adaptation loop stage.

D. **Calibration of the auditory model**
In this appendix the procedure we followed to "calibrate" the auditory (PEMO) model used in Chapters 4 and 5 is described.

E. **Other approaches for the memory template**
This appendix contains a description of the different template approaches that were tested in the simulations of Chapter 4. The use of these approaches did not lead to a satisfactory explanation of the experimental results.

A | Auditory frequency scales

This appendix contains a brief summary of two auditory frequency scales that are inspired by the concept of critical bands. These scales are (1) the critical-band rate z expressed in Bark, which is used in Chapter 2 and in Appendix B, and (2) the ERB scale expressed in ERB numbers (ERB_N), which is used in the remaining chapters. The purpose of this appendix is to provide a general understanding of both scales, the range of their values and their mapping to the frequency scale in Hz. A detailed comparison between both scales is not provided.

It is well known that the human hearing system acts as a frequency analyser, where different frequencies of the incoming signals stimulate different points of the basilar membrane in the inner ear. This frequency-to-position mapping can be approximated by the following analytical expression (Greenwood, 1990)[1]:

$$x = 16.67 \cdot \log_{10}\left(\frac{f}{165.4} + 1\right) \tag{A.1}$$

where the frequency f is expressed in Hz and x represents the distance in mm from the apex to the point of stimulation along the basilar membrane. The basilar membrane extends from the base (near to the middle ear) to the apex (innermost end of the cochlea), having an average length of 35 mm. The logarithmic relationship between the frequency f and the position x is indicated by the square red markers in Figure A.1. In the figure, the two auditory scales are also plotted as a function of the position x, showing an approximate linear relationship. This may not be surprising because the auditory frequency scales have been derived to "divide the frequency spectrum into bands of equal effectiveness" (Zwicker et al., 1957) and the relative width of such bands happened to be approximately constant around the point of excitation x. The auditory

[1] Equation A.1 can be obtained by replacing the constants $A = 165.4$ and $k = 1$ in Equation 1 of the study by Greenwood (1990).

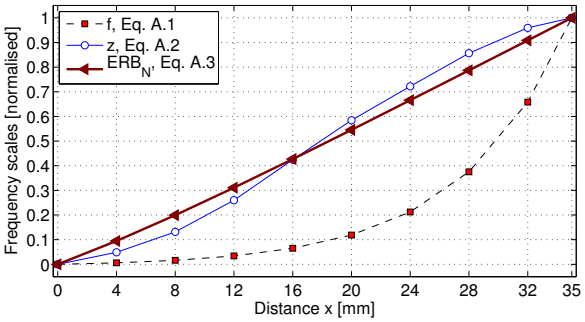


Figure A.1: Frequency-to-position mapping between different frequency scales (normalised between 0 and 1) and the corresponding point of stimulation x along the cochlea. In this figure, the positions x in mm were converted to frequency in Hz (Equation A.1). Subsequently the critical-band rate z and the ERB_N values were obtained using Equations A.2 and A.3, respectively. Both auditory scales have a nearly linear relationship with the point x.

scales differ, however, in the way they were derived. The critical-band rate scale z was derived by measuring the width of the "effective bands" in a number of experiments including detection thresholds with complex tones and narrow-band noises, amplitude and frequency modulation detection, localisation performance and loudness summation (Zwicker et al., 1957; Fastl & Zwicker, 2007, their Chapter 6). The ERB scale measures that bandwidth using a tone-in-notched-noise experiment (see, e.g., Patterson, 1976).

Both auditory scales are described next by providing an analytical expression that maps f onto the corresponding auditory scale. This appendix ends by providing a list of tabulated frequencies in Hz and their corresponding auditory frequencies z in Bark and in ERB_N.

A.1 Critical-band rate

An analytical expression to relate the frequencies z in Bark and f in Hz is given by Equation A.2 (Zwicker & Terhardt, 1980):

$$z = 13 \cdot \arctan\left(0.76 \cdot 10^{-4} f\right) + 3.5 \cdot \arctan\left(\left[\frac{f}{7500}\right]^2\right) \qquad \text{(A.2)}$$

This expression provides a close mapping between f in Hz and the critical-band rates z reported by Zwicker (1961). The bandwidth of each critical-band is 1 Bark. This leads to about 24 bands in the audible frequency range. The reader is referred to Zwicker et al. (1957) and Zwicker (1961) for further details about the critical-band rate scale.

Table A.1: List of frequencies in Hz and their mapping to the ERB-rate and the critical-band rate scales. The frequencies in ERB_N and Bark can be obtained using Equations A.3 and A.2, respectively.

Frequency f_c			Frequency f_c			Frequency f_c			Frequency f_c		
Hz	ERB_N	Bark	Hz	ERB_N	Bark	Hz	ERB_N	Bark	Hz	ERB_N	Bark
87	3.0	0.9	520	11.0	4.9	1547	19.0	11.4	3983	27.0	17.2
101	3.4	1.0	554	11.4	5.2	1628	19.4	11.7	4174	27.4	17.5
123	4.0	1.2	605	12.0	5.6	1749	20.0	12.2	4463	28.0	17.9
139	4.4	1.4	643	12.4	5.9	1839	20.4	12.6	4676	28.4	18.2
163	5.0	1.6	700	13.0	6.4	1975	21.0	13.0	4997	29.0	18.5
181	5.4	1.8	743	13.4	6.7	2075	21.4	13.3	5235	29.4	18.8
208	6.0	2.0	806	14.0	7.2	2226	22.0	13.8	5593	30.0	19.2
228	6.4	2.2	853	14.4	7.5	2338	22.4	14.1	5857	30.4	19.5
257	7.0	2.5	924	15.0	8.0	2506	23.0	14.5	6257	31.0	19.9
280	7.4	2.7	977	15.4	8.4	2630	23.4	14.8	6551	31.4	20.1
313	8.0	3.0	1056	16.0	8.9	2818	24.0	15.2	6996	32.0	20.5
338	8.4	3.3	1114	16.4	9.2	2956	24.4	15.5	7324	32.4	20.8
375	9.0	3.6	1202	17.0	9.7	3165	25.0	15.9	7819	33.0	21.1
402	9.4	3.9	1267	17.4	10.1	3319	25.4	16.2	9271	34.5	22.0
443	10.0	4.2	1365	18.0	10.6	3552	26.0	16.6	11581	36.5	23.0
474	10.4	4.5	1438	18.4	10.9	3723	26.4	16.8	15550	39.2	24.0

A.2 Equivalent rectangular bandwidth

The analytical expression that converts the frequency f in Hz to frequencies expressed in ERB_N is (Glasberg & Moore, 1990):

$$\text{ERB}_N = 9.2645 \cdot \ln\left(1 + 0.00437 \cdot f\right) \tag{A.3}$$

The use of a tone-in-notched-noise experiment to derive the bandwidth of a critical-band is believed to reduce off-frequency listening. This lead to a higher number of ERB bands with respect to the Bark scale with 39 bands up to the range reported in Table A.1. The reader is referred to Moore and Glasberg (1983) and Glasberg and Moore (1990) for further details about the ERB rate scale.

B | Modelling the sensation of fluctuation strength[1]

The sensation of fluctuation strength (FS) is elicited by slow modulations of a sound, either in amplitude or frequency (typically < 20 Hz), and is related to the perception of rhythm. In speech, such periodicities convey valuable information for intelligibility (prosody). In western music, most of the envelope periodicities are also found in that range. These are evidences of the potential relevance of FS in the perception of speech and music. In this appendix we present a model of fluctuation strength. Our model was developed taking advantage of the physical similarity between FS and the sensation of roughness. The FS model was then adjusted and fitted to existing experimental data collected using artificial stimuli, namely, amplitude- (AM) and frequency- (FM) modulated tones and AM broadband noise (BBN). The test battery of sounds also consists of samples of male and female speech and some musical instrument sounds. This FS model has been used in Chapter 2 of this thesis.

B.1 Introduction

Temporal fluctuations in amplitude and in frequency are found naturally in everyday sounds. Amplitude modulations (AM) are related to the envelope of a waveform, while frequency modulations (FM) to its fine structure. Envelope refers to the perceived acoustic amplitude of a sound that is integrated by the hearing system due to its slow response (or "sluggishness") to high rate (sound pressure) variations of its waveform. Two examples of everyday sounds are speech and music. Speech was described by Rosen (1992) as temporal fluctuating patterns with three partitions: envelope, periodicity and fine structure. The envelope contributes to, among other factors, prosody (i.e., duration, speech

[1]This chapter is based on:
R. García. (2015)"Modelling the sensation of fluctuation strength". M.Sc. thesis, Eindhoven University of Technology.
A. Osses, R. García, and A. Kohlrausch (2016). "Modelling the sensation of fluctuation strength". Proc. Mtgs. Acoust., 28(50005), pp. 1–8.

rhythm) and articulation, periodicity to intonation and fine structure to the timbre of a sound. With these concepts, it seems logical to assume that the characterisation of speech as temporal fluctuating pattern is also applicable to music. The link between prosody and Western music found by Patel, Iversen, and Rosenberg (2006) supports this assumption.

Two of the well-known classical psychoacoustical metrics are related to the perception of modulated sounds: fluctuation strength (FS) (Fastl, 1982, 1983) and roughness (Aures, 1985), for sounds modulated at slower frequencies (<20 Hz) and more rapid modulation rates (20-300 Hz), respectively. Both sensations show a bandpass characteristic with peaks at 4 Hz for FS and 70 Hz for roughness. The range of modulations below 20 Hz has been shown to be of special interest for speech intelligibility (Drullman et al., 1994; Shannon et al., 1995) as well as for the perception of rhythm, which is related to the average syllable rate at AMs of around 4 Hz (Leong et al., 2014).

Fluctuation strength is an attribute related to the perception of modulation in the range that we indicated as relevant for speech intelligibility (and potentially also for music). Roughness, however, is an attribute related to timbre (due to the higher modulation frequency range) that has taken more attention for its accepted influence in the perception of unpleasantness of a sound. There are, therefore, a number of published roughness models (e.g., Aures, 1985; Daniel & Weber, 1997; Kohlrausch et al., 2005). There is either less information about the algorithms to assess FS[2], or there are solutions that apply for a specific type of stimuli have been described (e.g., the FS model for AM tones and AM BBN, Fastl, 1982; Fastl & Zwicker, 2007). In this chapter a model of FS is presented. The similarities between FS and roughness listed above motivated the development of our implementation based on an existing roughness model (Daniel & Weber, 1997; García, 2015). There are, to our knowledge, two studies where a similar approach has been adopted (Zhou et al., 2015; Sontacchi, 1998)[3]. In comparison with those studies, the database of sounds used for developing and testing our algorithm is more diverse, including not only artificial sounds (AM and FM tones and AM BBN) but also a few cases of male and female speech and mu-

[2]The following commercial software packages include implementations of an FS algorithm: Pulse by Brüel & Kjær, ArtemiS by Head Acoustics GmbH, PAK by Müller-BBM, PAAS (Sontacchi, 1998). Technical aspects about their implementation and/or validation are not publicly available.

[3]The FS model by Zhou et al. has been developed in parallel to the model described in this appendix. Their model has been integrated into the AARAE toolbox for MATLAB http://www.densilcabrera.com/wordpress/aarae-2/ (last accessed on 18/07/2018).

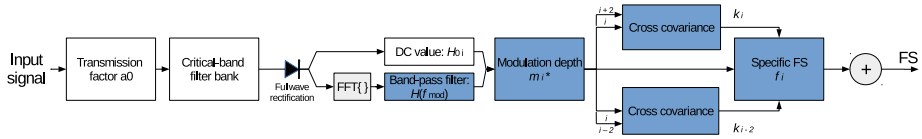


Figure B.1: Structure of our model of fluctuation strength.

sic samples, which were taken from the test battery of sounds used by Schlittmeier, Weissgerber, Kerber, Fastl, and Hellbrück (2012).

B.2 Description of the model

The algorithm used in our model of fluctuation strength (FS) was adapted from the roughness extraction algorithm described by Aures (1985) and Daniel and Weber (1997). The structure of the model is shown in Figure B.1, where the highlighted blocks represent the processing stages that we modified in our FS model. The model assumes that the total FS is the sum of partial contributions from N auditory filters and it is based on the concept of modulation:

$$\mathrm{FS} = \sum_{i=1}^{N} f_i = C_{FS} \cdot \sum_{i=1}^{N} (m_i^*)^{p_m} \cdot |k_{i-2} \cdot k_i|^{p_k} \cdot (g(z_i))^{p_g} \tag{B.1}$$

where N is the number of auditory filters (here $N = 47$), m^* is a generalised modulation depth, k refers to the normalised cross covariance between different auditory filters and $g(z_i)$ is an additional free parameter to introduce a weighting as a function of centre frequency. Frequencies equal or below 13 Bark[4] (1975 Hz) are unchanged and an attenuation (gain < 1) is applied to higher frequencies. The linear gains decrease monotonically from 1 (13 Bark or below) to 0.9, 0.7 down to 0.5, at 15.0 Bark (2730 Hz), 17.5 Bark (4174 Hz), and 23.5 Bark (13169 Hz), respectively. The product of all the elements in Equation B.1 as a function of the critical band i defines the specific fluctuation strength f_i. The parameters C_{FS}, p_m, p_k and p_g are constants optimised to fit the model. The values found for these parameters are $C_{FS} = 0.2490$, $p_m = 1.7$, $p_k = 1.7$ and $p_g = 1.7$.

In general, the model provides FS estimates for successive analysis frames. The frames have a duration of 2 s and a 90%-overlap and are

[4]The critical-band rate z expressed in Barks corresponds to one of the frequency scales that is inspired by the frequency representation in the auditory system. A brief overview of this scale is given in Appendix A.

gated on and off with 50-ms raised-cosine ramps. Each analysis frame is independently and successively passed through the processing blocks described below. For this reason from hereafter we refer to all analysis frames as the "input signal".

B.2.1 Spectral weighting: transmission factor a_0

To approximate the incoming signal to what arrives to the oval window (beginning of the inner ear), the transmission factor a_0 is applied. This factor introduces a frequency dependent gain that accounts for the sound transmission from free-field through the outer and middle ear. In our model a_0 was implemented as a 4096^{th}-order FIR filter.

B.2.2 Critical-band filter bank

In the frequency domain (N-point fast Fourier transform (FFT), frequency resolution $\Delta f = 0.5$ Hz), all frequency bins with amplitudes above the absolute hearing threshold are transformed into a triangular excitation pattern (Terhardt, 1979). The triangular excitation pattern produced by the frequency component f (in Hz) at a level L (in dB) has a constant lower slope S_1 of 27 dB/Bark and higher slope S_2 defined by:

$$S_2 = 24 + \frac{230}{f} - 0.2L \qquad \text{[dB/Bark]} \tag{B.2}$$

The slopes S_1 and S_2 are defined in the frequency domain using the critical-band rate scale. An analytical expression to relate the frequencies z in Bark and f in Hz is given by Equation A.2 in Appendix A.

The excitation patterns are a way to determine the contribution of a given component with frequency f_k (and level L_k) to another auditory filter, located at an "observation point" i, with a Bark distance of Δz Bark (keeping the same phase of the component at k). That contribution, $L_{k,i}$, can be expressed as:

$$\begin{aligned} L_{k,i} &= L_k - S_2 \Delta z = L_k - S_2(z_i - z_k) \qquad \text{if } f_k < f_i \\ L_{k,i} &= L_k - S_1 \Delta z = L_k - S_1(z_k - z_i) \qquad \text{if } f_k > f_i \end{aligned} \tag{B.3}$$

where z_i and z_k are the corresponding frequencies f_i and f_k in the critical-band rate scale that can be calculated using Equation A.2.

If we now consider 47 equally spaced "observation points" (with a spacing of 0.5 Bark) related to the frequency range from 0.5 Bark (50 Hz)

to 23.5 Bark (13169 Hz) and evaluate the individual contribution of each computed excitation pattern, 47 output (audio) signals are obtained. These 47 signals can be interpreted as the output of a critical-band filter bank with centre frequencies $z_i = 0.5 \cdot i$ Bark and bandwidth of 1 Bark. At the end of this stage each spectrum is converted back to the time domain using an inverse fast Fourier transform (IFFT), obtaining 47 $e_i(t)$ signals.

B.2.3 Generalised modulation depth $\mathbf{m}_i^*$

Each of the 47 signals $e_i(t)$ obtained from the critical-band filter bank is used to obtain an estimate of the modulation depth m^*. The so-called generalised modulation depth is calculated by dividing the root mean square (RMS) value of the weighted envelopes of $h_{BP,i}(t)$ by their DC values $h_{0,i}$. The DC value is calculated from the full-wave rectified time signals:

$$h_{0,i} = \overline{|e_i(t)|} \tag{B.4}$$

The weighted excitation envelopes are determined by:

$$h_{BP,i}(t) = \mathrm{IFFT}\{H(f_{\mathrm{mod}}) \cdot \mathrm{FFT}(|e_i(t)|)\} \tag{B.5}$$

The weighting function H is used because the fluctuations of the envelope are contained in the low part of the excitation patterns e_i in the frequency domain. The shape of the $H(f_{\mathrm{mod}})$ function was chosen to account for the bandpass characteristic of the FS sensation (with maximum at a modulation frequency f_{mod} of 4 Hz). The resulting $H(f_{\mathrm{mod}})$ was implemented as an IIR filter with passband between 3.1 and 12 Hz.

The RMS of the weighted functions $\overline{h_{BP,i}}$ is then used to obtain the generalised modulation depths:

$$m_i^* = \frac{\overline{h_{BP,i}}}{h_{0,i}} \tag{B.6}$$

In the original (roughness) model this ratio was limited to a maximum value of 1. FM tones represent a case where this limitation was often being applied, but their roughness in asper reaches larger values (3.2 asper for a 1.6-kHz tone, f_{mod} at 80 Hz, f_{dev} of ±800 Hz and 60 dB SPL) than those for FS in vacil (1.4-kHz tone, f_{mod} at 4 Hz, f_{dev} of ±700 Hz and 60 dB SPL). In our FS model we suggest to introduce a compression stage to the ratio m_i^* rather than a limitation. A compression ratio of

3:1 is applied when the modulation depth estimate exceeds a threshold of 0.7 units. This means that if m_i^* is 0.15 units above the threshold, i.e., $m_{i\,\text{input}}^* = 0.85$ the resulting modulation depth will be 0.05 (0.15/3) above threshold resulting in $m_{i\,\text{output}}^* = 0.75$.

B.2.4 Normalised cross covariance

In a discrete time domain the normalised cross covariance (in short, cross covariance) between the functions x and y, both being N samples long, is defined by Equation B.7 (see, e.g., van de Par & Kohlrausch, 1995, their Equation 2):

$$k = \frac{\sum xy - \frac{1}{N}\sum x \sum y}{\sqrt{\left[\sum x^2 - \frac{1}{N}(\sum x)^2\right]\left[\sum y^2 - \frac{1}{N}(\sum y)^2\right]}} \tag{B.7}$$

Within our computational model the cross covariance between adjacent critical bands is assessed to determine whether their modulations are in or out of phase. The more in-phase the modulations are determines to what extent the specific FS can be summed up to obtain the total FS. In this manner, the cross covariance between the channel i and the channels one Bark below $i-2$ and above $i+2$ are computed. In other words, to obtain the factor k_{i-2}, x and y in Equation B.7 have to be replaced by $h_{BP,i-2}$ and $h_{BP,i}$, respectively. Likewise, to obtain the factor k_i, x and y have to be replaced by $h_{BP,i}$ and $h_{BP,i+2}$.

B.3 Validation of the model

In order to fit and validate the FS model presented in this appendix, a set of stimuli with known values were chosen. Part of the set corresponded to artificial stimuli: AM tones, FM tones, and AM BBN. The rest of the stimuli were chosen from a set of everyday sounds. The reference sound to which an FS of 1 vacil is ascribed is an AM sine tone centred at $f_c = 1000$ Hz, modulated at an f_{mod} of 4 Hz and level of 60 dB. A summary of the artificial stimuli used in the validation is shown in Table B.1. For this set of stimuli, FS values obtained in perceptual experiments are available (Fastl & Zwicker, 2007, their Chapter 11). Additionally, a set of everyday sounds were extracted from the database of sounds used by Schlittmeier et al. (2012). That database consists of 70 sounds, out of which 7 representative sound samples were chosen. The selection of the samples was as follows: (a) three representative speech samples (one male voice, one female voice, babble noise); (b)

Table B.1: Artificial stimuli used to validate the FS model. The FS values were taken from Fastl and Zwicker (2007, their Chapter 11).

Type	Fixed parameters	SPL [dB]	Variable parameters (FS)
AM tone (reference)	$f_c = 1000$ Hz $m_{\text{index}} = 1$	60	$f_{\text{mod}} = \{4.00\}$ Hz (1.00) vacil
AM tone	$f_c = 1000$ Hz $m_{\text{index}} = 1$	70	$f_{\text{mod}} = \{1.00, 2.00, 4.00, 8.00, 16.0, 32.0\}$ Hz $(0.39, 0.84, 1.25, 1.30, 0.36, 0.06)$ vacil
FM tone	$f_c = 1500$ Hz $f_{\text{dev}} = \pm 700$ Hz	70	$f_{\text{mod}} = \{1.00, 2.00, 4.00, 8.00, 16.0, 32.0\}$ Hz $(0.85, 1.17, 2.00, 0.70, 0.27, 0.02)$ vacil
AM BBN	BW= 16000 Hz $m_{\text{index}} = 1$	60	$f_{\text{mod}} = \{1.00, 2.00, 4.00, 8.00, 16.0, 32.0\}$ Hz $(1.12, 1.58, 1.80, 1.57, 0.48, 0.14)$ vacil

Table B.2: Everyday sounds used to validate the FS model. An artificial noise (pink noise, Track Nr. 61) was also included. The average sound pressure level (SPL) of each sound is shown. For the changing-state speech samples and the ducks' quaking samples the maximum levels are also shown. The FS values were taken from Schlittmeier et al. (2012).

Type	Track Nr. / description	SPL [dB] L_{eq} (L_{max})	Reported FS [vacil]
Speech	1 / Narration, female voice	56.1 (67.2)	1.11
Speech	2 / Narration, male voice	60.0 (69.4)	1.21
Speech	23 / Eight talker babble noise	63.6 (67.8)	0.38
Music	24 / Strings concert	62.1	0.21
Music	31 / Violin solo	58.2	0.56
Animal	34 / Ducks' quaking	64.5 (73.4)	1.77
Noise*	61 / Broadband (pink) continuous noise	60.1	0.02

two music samples of soloist and ensemble playing, and (c) the sounds having minimum and maximum FS. For that database, Schlittmeier et al. (2012) used a commercial software to obtain their FS values. The selected samples are summarised in Table B.2.

B.3.1 Results for artificial stimuli

The artificial stimuli were used to fit the free parameters of the model: the constant C_{FS}, the BPF $H(f_{\text{mod}})$ and the exponents p_m and p_k. A value C_{FS} of 0.2490 was obtained. The $H(f_{\text{mod}})$ filter was fitted using 1000-Hz AM tones (with $1 \leq f_{\text{mod}} \leq 32$ Hz). As a result two cascaded IIR filters (4^{th}-order LPF and 2^{nd}-order HPF) producing a BPF between 3.1 and 12 Hz were obtained. The results of the FS model applied to the artificial sounds of Table B.1 are shown in Figure B.2. The model predicts qualitatively the fluctuation strength for AM tones, FM tones and AM BBN. There is, however, an overestimation of the FS estimates for FM tones especially for $f_{\text{mod}} > 4$ Hz (middle panel of the figure).

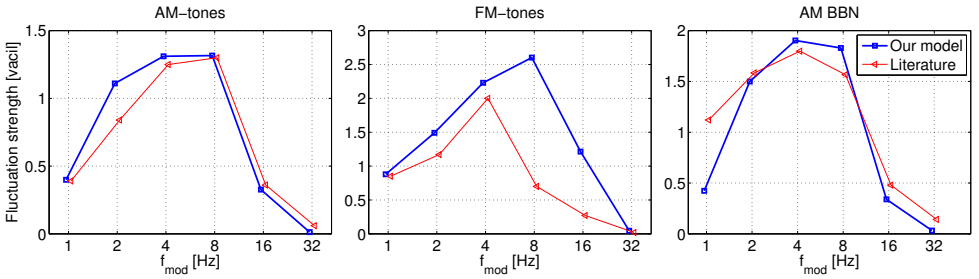


Figure B.2: Results obtained from the FS model for: (left panel) AM tones; (middle panel) FM tones, and (right panel) AM BBN.

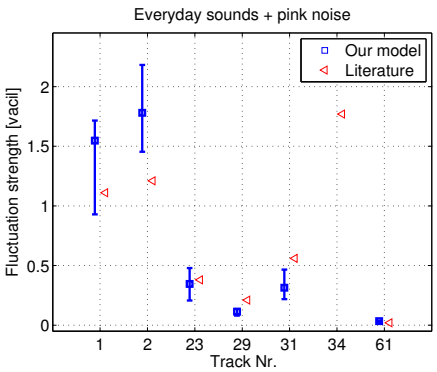


Figure B.3: Results obtained from the FS model using the everyday sounds detailed in Table B.2. The square markers correspond to median FS values along the sample duration. The errorbars represent the minimum and maximum FS. A high FS value (4.2 vacil) was found for track 34 (Ducks' quacking, not shown in the figure).

B.3.2 Results for everyday sounds

The FS values given by the model for the everyday sounds (and pink noise) of Table B.2 are shown in Figure B.3. For speech samples (Tracks 1 and 2) the median FS values were higher than the reference values by 0.45 and 0.58 vacil. For the eight-talker babble noise (Track 23), string concert (Track 29) and the pink noise (Track 61), the FS estimates seem to be in line with the reference values. For the violin solo (Track 31) there is an underestimation of the FS estimate (difference of 0.25 vacil). The highest FS estimate was found for the ducks' quacking (FS of 4.2 vacil).

B.4 Discussion

B.4.1 Artificial stimuli

Within the subset of artificial stimuli there is a reasonable agreement between the FS model and the experimental data for AM tones and

AM BBN noises. The model provides, however, overestimated FS values for FM tones with modulation frequencies above 4 Hz ($f_{\mathrm{mod}} > 4$ Hz), as shown in the middle panel of Figure B.2. Although the FS values show the expected band-pass characteristic as a function of f_{mod}, the maximum FS sensation is estimated to be at $f_{\mathrm{mod}} = 8$ Hz (instead of $f_{\mathrm{mod}} = 4$ Hz as for the experimental data). This shift in the maximum response of the band-pass characteristic is also observed in the roughness model (see Daniel & Weber, 1997, their Figure 9), where the maximum R estimate was found for an $f_{\mathrm{mod}} = 80$ Hz (instead of $f_{\mathrm{mod}} = 70$ Hz). It is known that when the FM comprises more than one critical band a higher FS sensation is elicited. With a carrier frequency of 1500 Hz (11.2 Bark) varied by a frequency deviation $\Delta f = \pm 700$ Hz more than 6 critical bands are covered (between 800 Hz or 7.1 Bark, and 2200 Hz or 13.7 Bark). To investigate the behaviour of the FS model for different frequency deviations, including deviations of less than one critical band, the following Δf values are tested: ± 25, ± 50 Hz (within one critical band), and ± 100, ± 200 Hz (more than one critical band). Sounds with a level of 72 dB SPL and carrier frequency $f_c = 851.8$ Hz (7.5 Bark) are conveniently chosen to allow a direct comparison of this new set of sinusoidally FM modulated tones with the hummer signals in acoustic mode 4 (see Chapter 2). The FS estimates for the new set of FM tones are shown in panel (a) of Figure B.4. For all tested frequency deviations, the FS estimates as a function of modulation frequency show a band-pass characteristic. The maximum FS estimates are 0.12, 0.35, 0.92, and 1.63 vacil for the FM tones with Δf deviations of ± 25, ± 50, ± 100, ± 200 Hz, respectively. Only for tones with Δf of ± 25 Hz the maximum estimate is found at $f_{\mathrm{mod}} = 4$ Hz, for the rest of the Δf values the maximum FS is found at $f_{\mathrm{mod}} = 8$ Hz. The patterns of specific fluctuation strength $\mathrm{FS}_{\mathrm{spec}}$ for the tones with $f_{\mathrm{mod}} = 4$ Hz are shown in panel (b) of the figure. As can be seen in the figure, the FS model returns $\mathrm{FS}_{\mathrm{spec}}$ patterns with significant contributions from critical bands that are not directly excited by the FM tones. For the FM tone with $\Delta f = \pm 200$ Hz and $f_{\mathrm{mod}} = 4$ Hz that has an FS of 1.33 vacil only 0.09 vacil are found in "on-frequency" critical bands (frequencies in the range 851.8 ± 200 Hz, i.e., between 6 and 8.8 Bark). In this example, the total off-frequency contribution is 1.24 vacil, with 0.26 vacil for frequencies below 6 Bark and 0.98 vacil above 9 Bark. This asymmetric contribution is, at least partly, due to the shallower slope of the critical-band filter bank towards higher frequencies (see Equation B.2). Although there is a lack of experimental

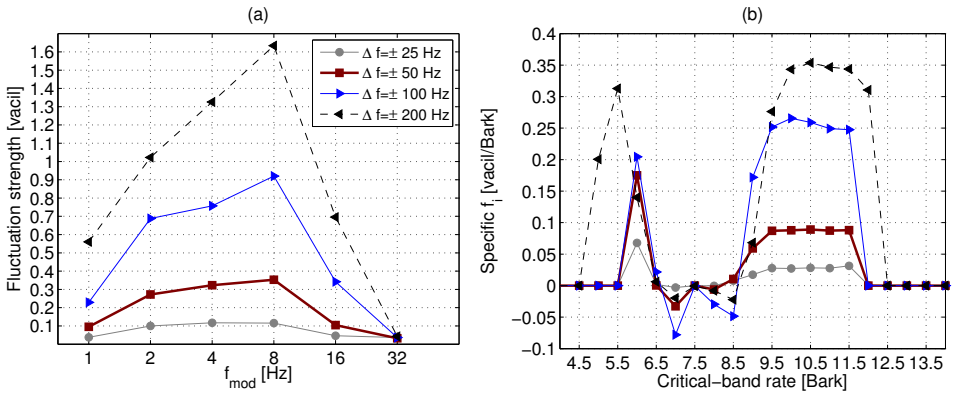


Figure B.4: (a) Fluctuation strength FS values and (b) specific fluctuation strength patterns FS_{spec} (only for tones with $f_{mod} = 4$ Hz) for sinusoidally FM tones centred at 851.8 Hz with a level of 72 dB SPL that are modulated using frequency deviations Δf of ±25, ±50, ±100, ±200 Hz. For this carrier frequency, the first two Δf values produce oscillations in frequency within one critical band (between 7 and 8 Bark). The FM tone with Δf of ±200 Hz covers the frequency range between 651.8 Hz (6 Bark) and 1051.8 Hz (8.8 Bark).

evidence for the FS estimates shown in panel (a) of Figure B.4, the band-pass characteristic built from experimental FS data collected by García from 20 participants for 70-dB FM tones, $f_c = 1500$ Hz, $\Delta f = \pm 700$ Hz, and $0 \leq f_{mod} \leq 128$ Hz (similar stimuli as used in panel (b) of Figure B.2) had its maximum FS value at $f_{mod} = 8$ Hz (García, 2015, his Figure 5.5(b)).

B.4.2 Everyday sounds

Within the set of everyday sounds there is a good approximation between FS values and the estimates in the reference paper for the eight-talker babble noise, the string concert and the pink noise samples. Higher FS values for the male and female voices and the ducks' quacking sounds and a lower value for the violin sample. For the male, female and ducks' quacking sounds our model provides high modulation depth m^* values, with a median across bands of 0.81, 0.95, and 0.86, respectively. The median cross covariance k for the same samples are 0.50, 0.20, and 0.83. It is noteworthy that the modulation depth m^* values in our model are assessed with respect to the DC values h_0, independent of the level of h_0. This means that the higher FS estimates in our model may be a consequence of the high m^* values. However, it is also important to point out that the estimates presented in the reference paper were obtained from another FS algorithm and, therefore, it is unclear whether those FS values have been validated experimentally.

B.5 Further extension of the model

For a number of cases our FS model shows a reasonable agreement with FS estimates obtained either experimentally (Fastl, 1983; Fastl & Zwicker, 2007) or by using commercially available software (Schlittmeier et al., 2012, using the PAK software). With respect to the literature, our model provides an overestimation of the FS estimates for FM tones (panel (b) of Figure B.2), male and female speech sounds (Tracks 1, 2), and ducks' quacking sound (Track 34 in Figure B.3). It is unclear whether this overestimation can be confirmed with existing experimental data, especially for natural sounds. The natural sounds that have overestimated FS values (speech and ducks' quacking sounds) are broadband and have large modulation depths m^*. We recommend to evaluate the dependency of fluctuation strength on stimulus level for sounds with inherent modulations (in amplitude and/or frequency) and to check whether the generalised modulation depth m^*, as used in our model, is a suitable measure for the variability of those modulations.

C Auditory modelling: Properties of the adaptation loops

The adaptation loops are included in models of the effective processing of the auditory system. This stage simulates the adaptive properties of the auditory system (see, e.g., Westerman & Smith, 1984; Kohlrausch et al., 1992). These properties refer to changes in the gain of the system as a consequence of changes in the level of the input signal.

The adaptation loops were first described by Püschel (1988) and then adopted by Münkner (1993) and Dau et al. (1996a) in the first versions of the models of the effective processing. A block diagram of the adaptation loops stage is shown in Figure C.1. In this appendix an in-depth analysis of their inherent properties is presented.

C.1 Input signal for the characterisation of the adaptation loops

In general the input to the adaptation loops is a signal after band-pass filtering and inner-hair cell processing (Stages 2−4 of the PEMO model).

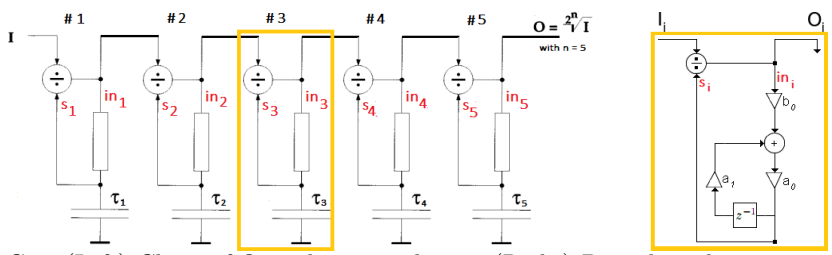


Figure C.1: (Left) Chain of five adaptation loops. (Right) Digital implementation of the adaptation loop i. The labels in_i indicate the input to the adaptation loop i, which in turn represents the input for the divisor of the next element. The input and output of the adaptation loop i are indicated by $\mathbf{I}_i$ and $\mathbf{O}_i$. We keep however the notation of in_i (which is equal to $\mathbf{I}_i$) and s_i ($\mathbf{O}_i[n] = \text{in}_i[n]/s_i[n-1]$) because the structure between in_i and s_i is an IIR LPF which is characterised by Equation C.2, whose constants are derived from τ_i.

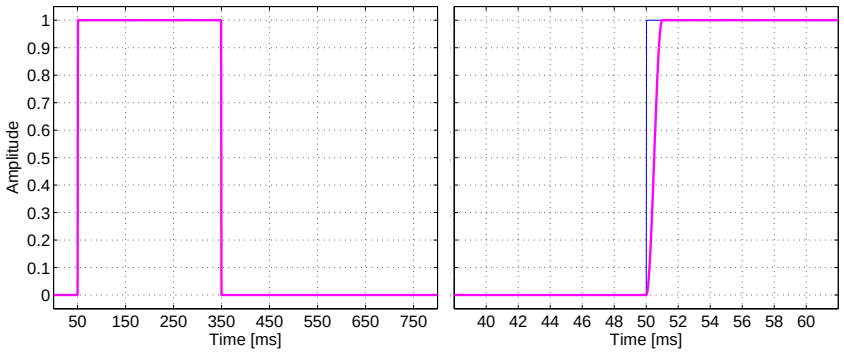


Figure C.2: Steady-state signal used to generate the analyses presented in this section. The signal has unit amplitude, duration of 300 ms with the signal onset at 50 ms and it includes an up-down cosine ramp of 1 ms. The cosine ramp introduces a similar effect to the pulse as the inner-hair cell stage of the PEMO model would do. The right panel corresponds to the same pulse as in the left panel but zoomed in to appreciate the raised cosine ramp.

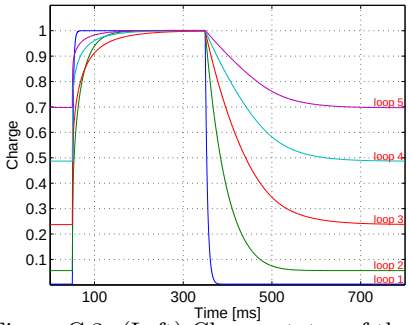
In the analyses of this appendix we only account for the inner-hair cell processing. Therefore, the input $x[n]$ corresponds to a digital waveform after half-wave rectification and a low-pass filtering with a cut-off frequency of 770 Hz. The input $x[n]$ is scaled between 0 and 1.

The analyses presented in the subsequent subsections (C.2 and C.3) are generated using the pulse signal that is shown in Figure C.2. The pulse signal has unit amplitude (steady-state input of 100 dB SPL), a duration of 300 ms and it is preceded and succeeded by 50 ms and 450 ms of silence, resulting in a signal 800 ms long. To facilitate the reproducibility of the analyses, the pulse was ramped up and down with a cosine ramp of 1 ms. The cosine ramp introduces a similar effect to the pulse signal as the stage of inner hair-cell processing of the PEMO model would do (Stages 3 and 4 in the diagram of Figure 4.1, page 56)[1].

C.2 Adaptation and use of the RC analogy

The adaptation stage comprises a chain of 5 adaptation loops, which is shown in Figure C.1. Each adaptation loop corresponds to a Resistor-Capacitor (RC) circuit that acts as a low-pass filter between the node in_i and the value s_i, with $i = 1, 2, 3, 4, 5$. The output s_i represents the charging state of the low-pass filter. The low-pass filters are implemented as first-order IIR filters and their time constants relate to their cut-off frequencies according to: $\tau_i = 1/(2\pi \cdot f_{\text{cut-off}})$. The outputs s_i for a steady-state input of amplitude 1 are shown in Figure C.3. As

[1] For this analysis the effect of the Stages 1 (Outer and middle ear filtering) and 2 (ERB filter bank) of the PEMO model were omitted.



Loop i	τ_n [ms]	Initial state $s_{0,i}$	IIR Coeff. ($f_s = 44100$ Hz) a_0	$a_{1,i}$	$b_{0,i} \cdot 10^{-3}$
1	5	0.0032	1	-0.99548	4.52490
2	50	0.0562	1	-0.99955	0.45341
3	129	0.2371	1	-0.99982	0.17577
4	253	0.4870	1	-0.99991	0.08962
5	500	0.6978	1	-0.99995	0.04535

Figure C.3: (Left) Charge status of the five adaptation loops when feeding the steady-state input shown in Figure C.2. The steady signal was preceded by the inner-hair cell processing, so the onset and the offset of the signal were slightly smoothed. (Right) Some parameters that characterise the adaptation loops. The steady-state charge was assessed considering a minimum instantaneous amplitude of $\mathrm{lvl}_{\mathrm{min}} = 1 \cdot 10^{-5}$ (0 dB SPL) of the input signal.

can be seen in the figure, the charge of each RC component is a value between the initial state of charge of the RC components and 1, and the shorter the time constant the faster the charge or discharge occur. An uncharged RC component amplifies the incoming signal. A fully charged RC component does not alter the amplitude of the incoming signal. This action produces rapid fluctuations (large amplitudes) while the RC components are being charged and slower fluctuating amplitudes when they are already charged. For any stationary input level I, i.e., when all RC components are charged, an output of $O = \sqrt{I}$ is obtained after the first adaptation loop. After $N = 5$ adaptation loops the output is $O = \sqrt[2^N]{I}$. This transformation provides approximately a logarithmic transformation as shown in panels A and C of Figure C.6 (see also Dau et al., 1996a, page 3617). As shown in the Table on the right of Figure C.3, this gives a stationary value of 0.6978 for an input of 0 dB SPL (minimum amplitude of $\mathrm{lvl}_{\mathrm{min}} = 1 \cdot 10^{-5}$). With this minimum input value each adaptation loop has initial conditions (initial-state levels $s_{0,i}$) given by:

$$s_{0,i} = \frac{1}{a_0} \cdot \sqrt[2^i]{\mathrm{lvl}_{\mathrm{min}}} \qquad \text{with } i = 1, 2, 3, 4, 5 \tag{C.1}$$

The difference equation that characterises the RC component in each adaptation loop i (between the input in_i and the output s_i, see Figure C.1) is given by:

$$a_0 \cdot s_i[n] - a_{1,i} \cdot s_i[n-1] = b_{0,i} \cdot \mathrm{in}_i[n] \tag{C.2}$$

The previous difference equation corresponds to a first-order IIR LPF. The coefficient a_0 is always unity. The coefficients $a_{1,i}$ and $b_{0,i}$ are ob-

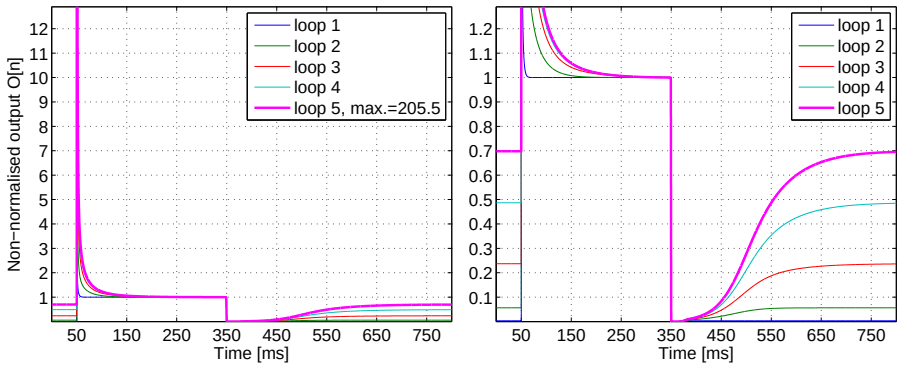


Figure C.4: Output of the adaptation loops for the steady-state input shown in Figure C.2. The maximum non-normalised output of the adaptation loops reach an amplitude of 206. In the right panel the ordinate has been zoomed in. The initial state of charge of the adaptation loops has an amplitude of 0.6978 and goes back to this value.

tained as:

$$a_{1,i} = \exp\left(-\frac{1}{\tau_i \cdot f_s}\right) \qquad b_{0,i} = 1 - a_{1,i} \tag{C.3}$$

The filter parameters are shown in the Table on the right of Figure C.3 for a sampling frequency $f_s = 44100$ Hz.

The output of the adaptation loops stage for our test pulse is shown in Figure C.4 (thick line in magenta). The thinner lines (not fully visible) correspond to the intermediate signals after loops 1 to 4.

C.3 Output of the adaptation stage

An appropriate scaling has to be applied to the output $O[n]$ of the adaptation loops stage that is shown in Figure C.4. As can be seen in the figure, the steady-state point of the curve is 1 (because the input pulse has an amplitude of 1) and the steady-state point of the curve after the signal offset corresponds to the steady-state value of the last adaptation loop (value of 0.6978). These amplitudes should be mapped in a way that a value $O[n] = 1$ is converted into $\Psi[n] = 100$ and a value $O[n] = 0.6978$ is converted into $\Psi[n] = 0$. The expression to obtain such a normalisation, expressed in MU, is given by:

$$\Psi[n] = 100 \cdot \frac{O[n] - 0.6978}{1 - 0.6978} \tag{C.4}$$

where $O[n]$ is the output of the last adaptation loop.

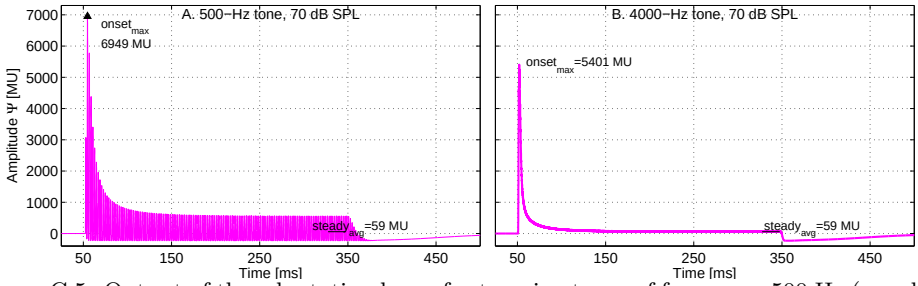


Figure C.5: Output of the adaptation loops for two sine tones of frequency 500 Hz (panel A) and 4000 Hz (panel B), level of 70 dB SPL including 2.5 ms raised cosine ramps. This figure is similar to (Breebaart et al., 2001, their Figure 2). The onset and steady-state amplitudes are 6949 and 59 [MU] for the 500-Hz tone and 5401 and 59 [MU] for the 4000-Hz tone, respectively.

Note that with this scaling the minimum possible value (during undershoot) is scaled to −230.9 MU (if a value $O[n] = 0$ is used in Equation C.4). A maximum value occurs when all the loops are at rest (initial state of charge) and a big change in the input amplitude is introduced. In our example with the artificial pulse signal, this generates a non-normalised amplitude of 205.5 which corresponds to an amplitude of 67605 MU.

In the next section a characterisation of the adaptation loops response to ramped pure tones is provided. The scaled amplitudes $\Psi[n]$ in MU are reported for the onset and steady-state responses of the tones as a function of their input level.

C.4 Input-output characteristic

In this section a set of pure tones is used to characterise the behaviour of the adaptation loops. The pure tones have centre frequencies of 500 Hz and 4000 Hz and a level that is varied from 0 to 100 dB SPL in steps of 10 dB. The output for two signals presented at a level of 70 dB SPL, duration of 300 ms, including 2.5 ms raised cosine ramps are shown in Figure C.5. For these signals the maximum amplitudes correspond to 6900 and 5400 MU for the 500-Hz and 4000-Hz tones, respectively. These values indicate that the adaptation loops produce a strong overshoot effect. This overshoot should be related to maximum firing rates for similar stimuli in the auditory nerve (Münkner, 1993; Dau et al., 1997a). In the study by Westerman and Smith (1984) similar stimuli were used to obtain neurophysiological measures of firing-rate patterns in the auditory nerve of the Mongolian gerbil. In their study they found that 40-dB pure tones generated an average firing rate of 642 spikes/s during the first

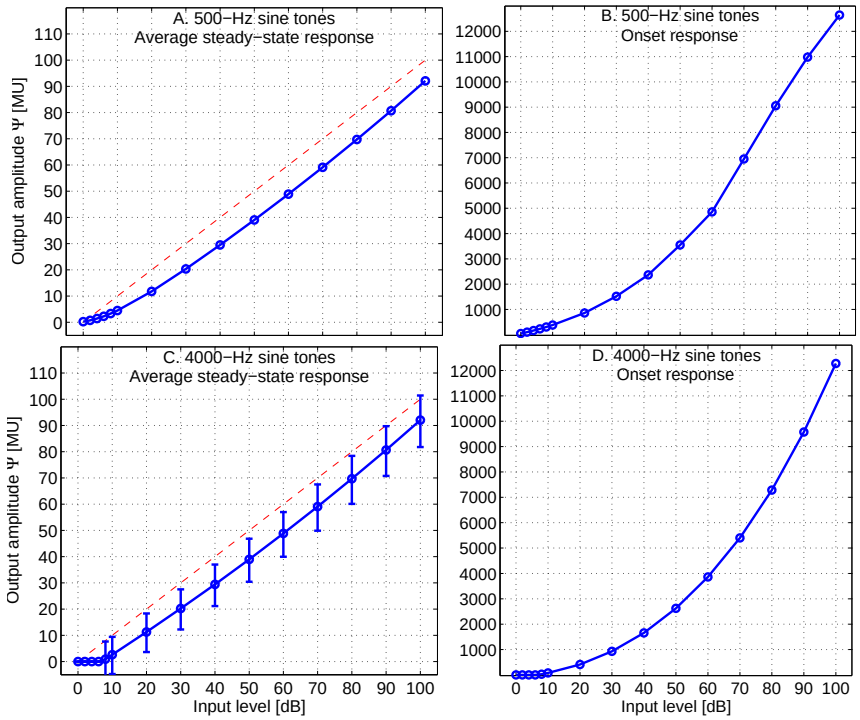


Figure C.6: Input-output characteristic function of the adaptation loops for 500-Hz (top panels) and 4000-Hz pure tones (bottom panels). The error bars in panel C indicate the minimum and maximum amplitude Ψ of the averaged amplitudes. They show that the fine structure of the 4000-Hz tone in its steady-state part (Figure C.5, panel B) has not been completely removed by the 770-Hz LPF. The error bars are not shown for the 500-Hz tone (panel A) but they would be very large since almost no fine structure is removed by the LPF.

20 ms of stimulation and an average of 107 spikes/s for the last 20 ms (driven-steady-state component). This represents a ratio of 6 between the rapid and steady averages.

The overshoot response of the adaptation loops as described so far reaches values of nearly 13000 MU for the 500-Hz sine tone at 100 dB SPL, as shown in panel B of Figure C.6. That overshoot has a ratio of more than 130 times the steady-state value of 92.1 MU.

In the next section a compression stage introduced to the output $O[n]$ of the adaptation loops is described. This compression was introduced by Münkner (1993) and adopted by Dau et al. (1997a) to limit the ratio between the onset response and the steady-state response of the adaptation loops.

The steady-state responses shown in Figure C.6 were obtained by

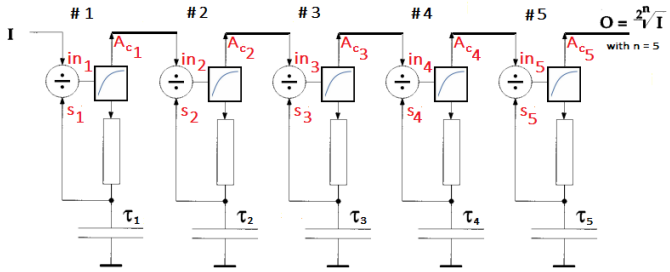


Figure C.7: Chain of five adaptation loops including logistic growth compressors to limit the overshoot response of the system.

averaging amplitudes in the last 20 ms of the internal representation of 300-ms long sine tones. The onset responses were obtained as the maximum of those amplitudes. We used similar integration periods as reported by Westerman and Smith (1984).

C.5 Overshoot limitation

This stage introduces a limitation to the overshoot response of the adaptation loops in a way that the maximum output values $\Psi[n]$ produce an amplitude comparable to the average firing rate at the level of the auditory nerve. The so called overshoot limitation is implemented as a compressor with a compression ratio that follows a logistic growth.

The following expression is used to limit the individual outputs of each adaptation loop (non-normalised outputs):

$$A_{c,i} = \begin{cases} \text{in}_i & \text{for } \text{in}_i \leq 1 \\ \frac{2 \cdot C_i}{1+\exp\left[\frac{-2}{C_i} \cdot (\text{in}_i - 1)\right]} - (C_i - 1) & \text{for } \text{in}_i > 1 \end{cases} \tag{C.5}$$

This equation implements a compression to the input in_i with output $A_{c,i}$. The compressor has a threshold of 1 and a limiter threshold $\mathsf{thres}_{\text{lim},i}$, that depends on the constant C_i. In turn, the constant C_i depends on the initial charge of each adaptation loop. The quantity $(\text{in}_i - 1)$ corresponds to the amount of exceedance above the non-normalised amplitude of 1. The block diagram of the adaptation loops including the compressive stage is shown in Figure C.7.

The constant C_i is obtained by defining an arbitrary limiter threshold $\mathsf{thres}_{\text{lim},i}$. A limiter factor limit has to be chosen. This factor is related to the actual limiter threshold $\mathsf{thres}_{\text{lim},i}$ according to Equation C.6:

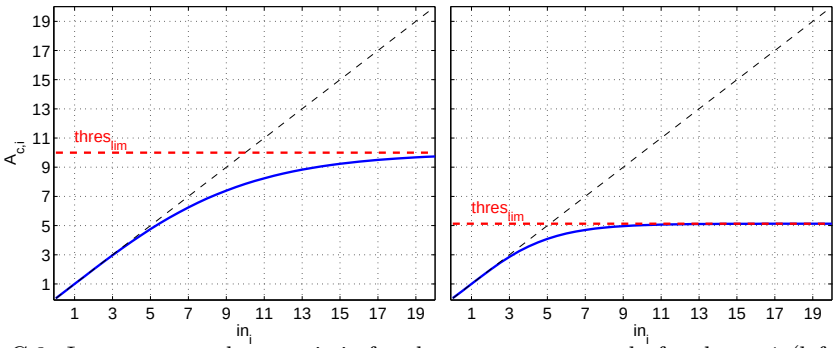


Figure C.8: Input-output characteristic for the compressors used after loops 1 (left panel) and 5 (right panel), when a limiter factor of 10 is used. With an initial status of charge $s_{0,1} = 0.0032$ the limiter threshold $\mathsf{thres}_{\mathsf{lim},1}$ turns to be 10 ($C_1 = 9$). For an initial status of $s_{0,5} = 0.6978$ the limiter threshold $\mathsf{thres}_{\mathsf{lim},5}$ turns to be 5.1 ($C_5 = 4.1$).

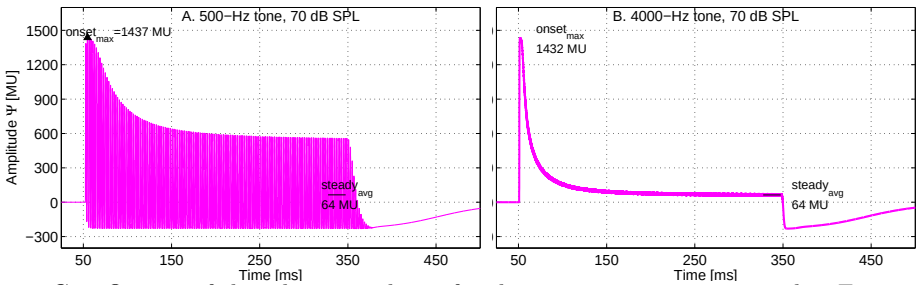


Figure C.9: Output of the adaptation loops for the same two sine tones used in Figure C.5 but for an overshoot limitation of 10. The onset and steady-state amplitudes are 1437 and 64 MU for the 500-Hz tone and 1432 and 64 MU for the 4000-Hz tone, respectively.

$$\begin{aligned} \mathrm{thres}_{\mathrm{lim},i} &= (1 - s_{0,i}^2) \cdot \mathrm{limit} \\ C &= \mathrm{thres}_{\mathrm{lim},i} - 1 \end{aligned} \tag{C.6}$$

This means that the higher the initial state of charge $s_{0,i}$ the lower the limiter threshold $\mathsf{thres}_{\mathsf{lim},i}$. The input-output characteristic function of the compressors used after loops 1 and 5 are shown in Figure C.8 for $\mathsf{limit}= 10$. This limiter factor has been adopted in almost every version of the auditory models where an overshoot limitation has been applied.

The effects of adopting an "overshoot limitation of 10", i.e., of using a limiter factor $\mathsf{limit}= 10$, on the two 70-dB pure tones used in Figure C.5 are shown in Figure C.9. The onset of the signals was reduced from 6949 to 1437 MU for the 500-Hz tone and from 5401 to 1432 MU for the 4000-Hz tone. The average steady-state response of the signals was slightly increased from 59 to 64 MU for both tones. Particularly for the 4000-Hz

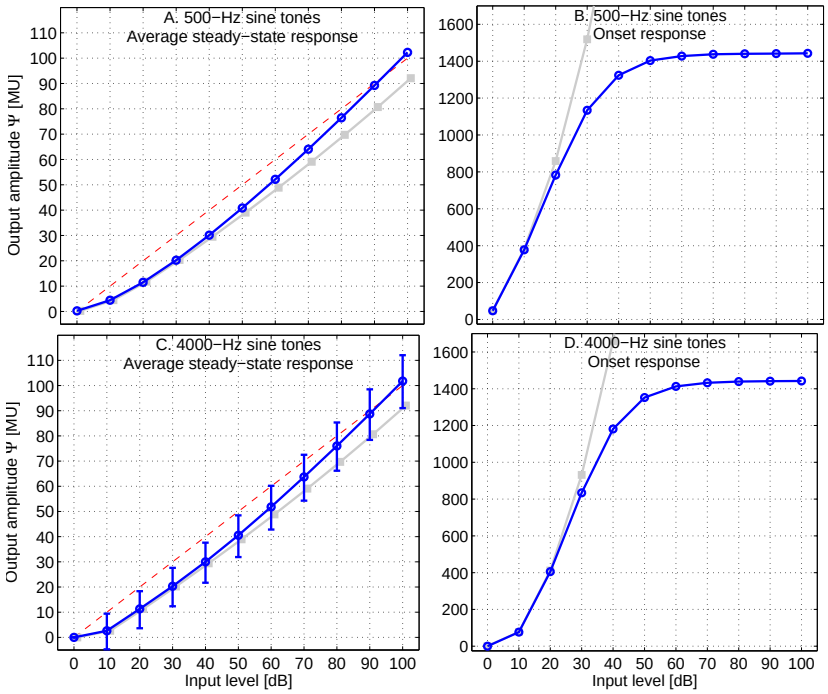


Figure C.10: Input-output characteristic function of the adaptation loops for 500-Hz pure tones (top panels) and 4000-Hz pure tones (bottom panels) using an overshoot limitation with a factor of 10. The error bars in panel C indicate the minimum and maximum amplitude Ψ of the averaged amplitudes. They show that the fine structure of the 4000-Hz tone in its steady-state part has not been completely removed by the 770-Hz LPF. The grey lines indicate the input-output functions when no overshoot limitation is used (as in Figure C.6).

tone, its steady-state response should remain unmodified since its amplitudes in the last 20 ms never go above the compression threshold of 1 (i.e., Ψ =100 MU). The slight increase in the average Ψ amplitudes is produced, however, by the fact that a lower $\Psi_{\max}$ introduces less compression to subsequent samples in the adaptation loops as a consequence of entering lower amplitudes to the divisor elements. This leads to a steady-state point that is reached somehow later in time in comparison to the situation where the adaptation loops are not limited.

The input-output characteristic functions for the steady-state and onset responses of the adaptation loops for limit= 10 are shown in Figure C.10. The steady-state response of the 4000-Hz tone is shown with error bars indicating the maximum and minimum Ψ amplitudes. This information is shown to point out that the fine structure of the 4000-Hz tone is not fully removed by the fifth-order 770-Hz LPF as could be assumed when inspecting the panel B of Figure C.9. From the right panels

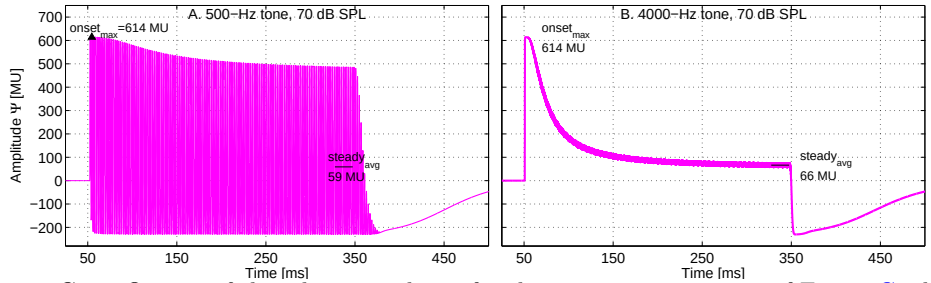


Figure C.11: Output of the adaptation loops for the same two sine tones of Figure C.5 but for an overshoot limitation of 5. The onset and steady-stage amplitudes are 614 and 59 MU for the 500-Hz tone and 614 and 66 MU for the 4000-Hz tone, respectively.

of Figure C.10 it can be seen that the onset responses are (1) almost not affected for input levels up to 20 dB, (2) compressed for levels between 20 and 50 dB, and (3) limited for levels above 50 dB.

The ratio between onset and steady-state responses is shown in Figure C.13. The ratio considering the limiter factor limit= 10 is indicated by the black lines in the figure. For tones of 50 dB or more, the ratio ranges from a factor of about 35 down to a factor of 15. The behaviour is similar in that level range for the tones of 500 and 4000 Hz. The ratio stays above the intended limitation of 10 times the steady-state level. In other words, with an overshoot limitation of 10, the adaptation loops are still overestimating the signal onsets compared with the neurophysiological findings of Westerman and Smith (1984).

For sounds with prominent onset characteristics, as it is the case for the piano sounds used in Chapters 3, 4, and 5, an overshoot limitation with a factor of 5 was adopted. The use of this new limiter factor limit= 5 is the key for the success of the simulations of perceptual similarity in this thesis. The effect of such a limitation for the 500 and 4000-Hz tones is shown in Figure C.11 and the underlying input-output characteristic functions are shown in Figure C.12. The ratio between onset and steady state response is indicated by the blue lines in Figure C.13. With the overshoot limitation of 5, the ratio stays below a factor of 15 for levels of 50 dB or more, and below 10 for levels above 65 dB (down to a factor of 5.8). These ratios are closer to the intended overshoot limitation described in the literature.

To conclude the revision of the properties of the adaptation loops stage, we wanted to point out one aspect about the use of low level input signals into the system. As can be seen in panels C and D of Figure C.11, the 4000-Hz tones need to have a level of at least 8 dB

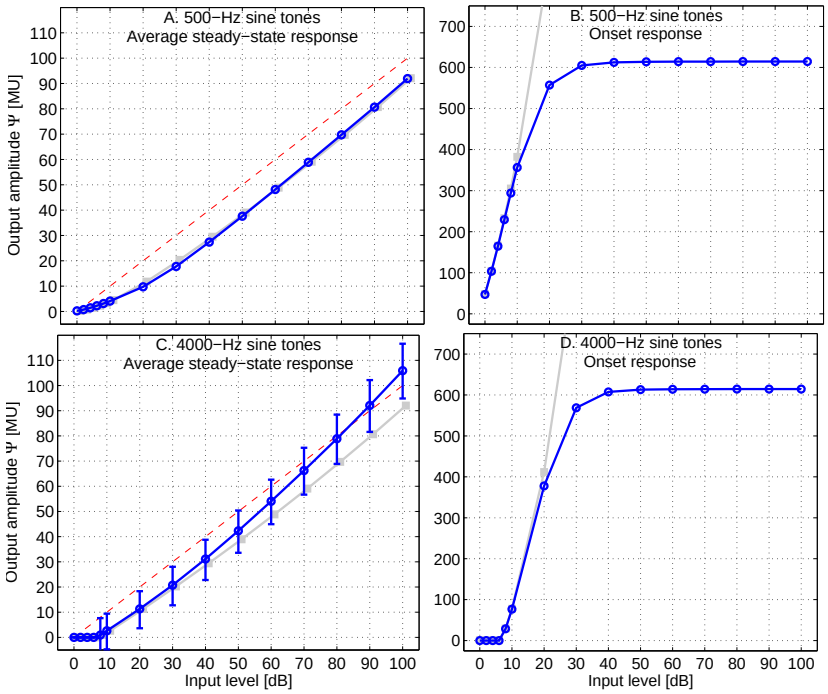


Figure C.12: Input-output characteristic function of the adaptation loops for 500-Hz pure tones (top panels) and 4000-Hz pure tones (bottom panels) using an overshoot limitation with a factor of 5. The error bars in panel C indicate the minimum and maximum amplitude Ψ of the averaged amplitudes. They show that the fine structure of the 4000-Hz tone in its steady-state part has not been completely removed by the 770-Hz LPF. The grey lines indicate the input-output functions when no overshoot limitation is used (as in Figure C.6).

SPL to generate a non-zero output. Although only instantaneous levels below 0 dB SPL are ignored (amplitudes below $1 \cdot 10^{-5}$), the tones are also subjected to the fine-structure removal (use of the 770-Hz LPF). From the input-output characteristic functions of the figure, it can be inferred that this processing introduces an attenuation between 6 and 8 dB for frequency components of 4000 Hz.

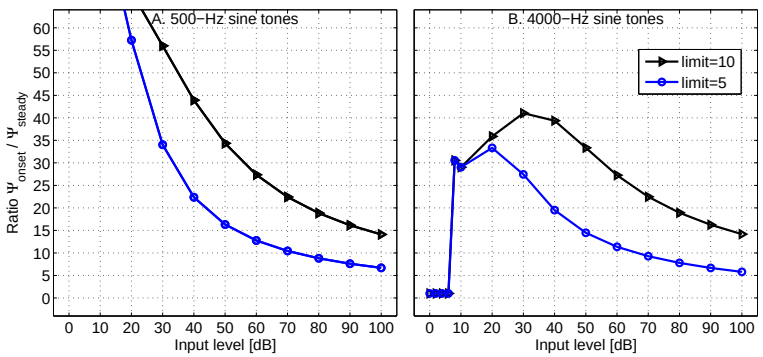


Figure C.13: Ratio between onset and steady responses for 500 (panel A) and 4000-Hz (panel B) tones for an overshoot limitation with factors of either 5 or 10. As discussed in the text, the 4000-Hz tones are more affected by the attenuation introduced by the fifth-order 770-Hz LPF, generating null-amplitude outputs for tones with levels below 8 dB SPL.

D | Auditory modelling: Calibration of the auditory model

In this appendix the procedure we followed to "calibrate" the auditory (PEMO) model used in Chapters 4 and 5 is described. The calibration consisted of finding a value for the variability σ of the internal (Gaussian) noise in a way that the performance of the artificial listener meets a given criterion. More specifically, every time a parameter in the auditory model was added, removed, or modified, the variability σ of the internal noise was adjusted (see Equation 4.5 in Chapter 4) to fulfil an intensity-discrimination task with a 70.7% score at a predefined test intensity.

In this appendix, two different σ values were used. A standard deviation $\sigma = 3.4$ MU was used to replicate simulation results of the PEMO model for the auditory tasks reported by Jepsen et al. (2008). A value of $\sigma = 10.1$ MU was used to limit the performance of the artificial listener to an intensity-discrimination task using piano sounds. The latter σ value was used to obtain the simulation results shown in Chapters 4 and 5. In this appendix we do not provide a critical analysis of how similar our simulation results using the PEMO model are with respect to the results presented by Jepsen et al. (2008). The objective was to replicate reported simulation results with the PEMO model as used in this thesis. The interested reader may directly compare our results to those presented by Jepsen et al. (2008).

D.1 Simulation procedure

All simulations reported in this appendix were run using the AFC toolbox for MATLAB (Ewert, 2013). In this toolbox an artificial listener was enabled to conduct the listening experiments presented in the subsequent sections. The artificial listener processed the incoming sounds using the auditory PEMO model. The experiments were all implemented as 3-AFC tasks using a two-down one-up tracking rule. Both the adjustable

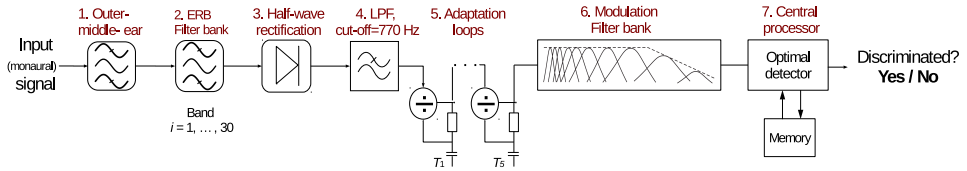


Figure D.1: Block diagram of the PEMO model. Refer to Chapter 4 for a detailed description of each of the model stages.

parameter and the suprathreshold level (used to derive the template in the auditory model) differed from task to task and are clearly indicated in the corresponding experimental description. Each simulated threshold was assessed 6 times. The median and IQR of the simulated thresholds based on those 6 estimates are reported.

D.2 Configuration of the auditory model

The block diagram of the PEMO model is shown in Figure D.1 (replotted from Figure 4.1). The final set of parameters used in our model simulations are listed in this section.

Stage 1. Outer and middle-ear: two cascaded 512-tap FIR filters that produce the combined frequency response shown in Figure 4.2.

Stage 2. Gammatone filter bank: set of 30 or 31 frequency bands with f_c between 80 and 8000 Hz, spaced at 1 ERB_N, as described by Hohmann (2002). Only the real part of the complex-valued outputs of the filter bank are used.

Stage 3 and 4. Half-wave rectification and LPF: the half-wave rectification is followed by a chain of five cascaded first-order IIR filters with $\mathrm{f_{cut\text{-}off}}$= 2000 Hz. The chain of filters produces a combined response that has an $\mathrm{f_{cut\text{-}off}}$ of 770 Hz.

Stage 5. Adaptation loops: the adaptation loops have time constants $\tau = 5, 50, 129, 253, 500$ ms. A limiter factor limit= 5 is used.

Stage 6. Modulation filter bank: the modulation filter bank we used is as reported by Jepsen et al. (2008).

Stage 7. Central processor: the decisions made by the model used all auditory channels (30 or 31 bands) with centre frequencies between 80 and 8000 Hz.

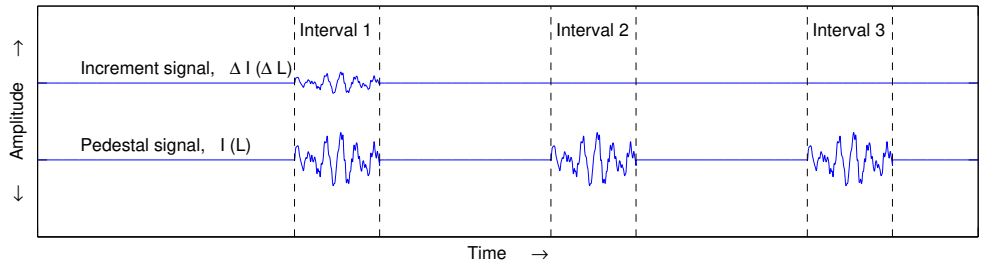


Figure D.2: Diagram of an increment-detection experiment implemented as a 3-AFC task where the first interval contains the target signal. In the course of an adaptive track the pedestal signal I stays at a constant level while the level of the increment signal I_i is adjusted using a two-down one-up rule. The increment signal is a scaled version of the pedestal signal, meaning that we simulate a coherent (in-phase) addition of the pedestal and increment signal. In this appendix we describe the intensity differences as JND values in level ΔL. In this way, for a pedestal level L of 60 dB, an increment signal L_i of 41.8 dB produces a total level L_t of 61 dB (i.e., a 1-dB increment).

D.3 Intensity discrimination

The discrimination of pure tones and broad-band noise is known to have JNDs in intensity (ΔI) that are approximately a constant fraction of their intensity I (Miller, 1947; Rabinowitz, 1970). We describe intensity differences as JNDs ΔL in level. A diagram of the experiment implemented as a 3-AFC task is shown in Figure D.2. The pedestal signal has a level L that is kept constant. The increment signal is a scaled (in-phase) version of the pedestal signal and it has a level L_i that produces a signal with a total level $L_t = 20 \cdot \log_{10}(10^{L/20} + 10^{L_i/20})$. The level difference ΔL produced by the increment level L_i is, therefore, $\Delta L = L_t - L$, expressed in dB.

D.3.1 Implementation as an adaptive procedure

For an implementation of the increment-detection task using an adaptive procedure it is convenient to express the increment level L_i as a level relative to the pedestal (test) level L. In this way, an increment level $L_{i\ \mathrm{rel}} = -18.2$ dB is a level 18.2 dB below the pedestal level L. A level $L_{i\ \mathrm{rel}} = -18.2$ dB produces a level difference ΔL of 1 dB. For $L = 60$ dB, a level $L_{i\ \mathrm{rel}} = -18.2$ dB corresponds to $L_i = 42.8$ dB, producing a total level $L_t = 61$ dB and therefore a $\Delta L = 1$ dB.

The parameters of the adaptive procedure used for three intensity-discrimination experiments –using pure-tones, broad-band noise, and piano sounds– were as follows:

- **Fixed parameter**: test (pedestal) levels L from 20 to 80 in steps of 10 dB (7

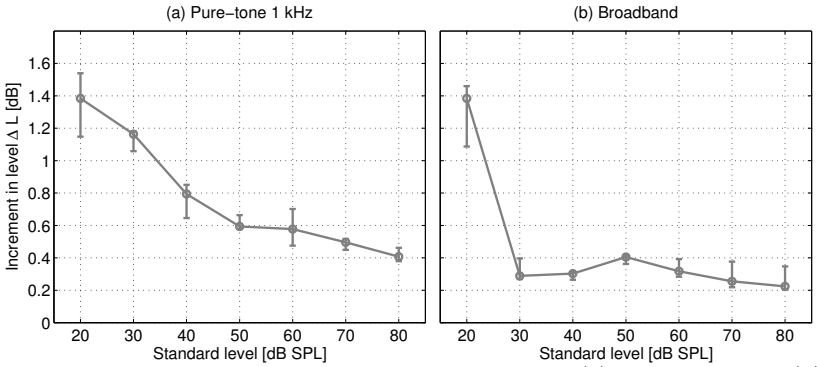


Figure D.3: Results of the intensity-discrimination task with (a) pure tones, and (b) broadband noise.

conditions = 7 adaptive procedures). For the intensity discrimination with anechoic piano sounds (as in Chapter 3) only one level L was tested for each piano (7 pianos = 7 adaptive procedures).

- **Adjustable parameter**: increment level $L_{i\ \mathrm{rel}}$.
- **Starting value**: $L_{i\ \mathrm{rel}} = -30$ dB ($\Delta L = 0.27$ dB)
- **Step size**: $L_{i\ \mathrm{rel}}$ was adjusted in steps of 4, 2, 1, and 0.5 dB, i.e., the step size was halved every two reversals until a step size of 0.5 dB was reached.
- **Number of reversals**: 12. The median of the last 6 reversals (at the step size of 0.5 dB) is used to estimate the JND in level (ΔL) for the corresponding test (pedestal) level.
- **Suprathreshold level**: $L_{i\ \mathrm{rel,supra}} = -5$ dB ($\Delta L = 3.9$ dB)

D.3.2 Evaluation of pure-tones and broad-band noise

(Obtained standard deviation of the internal noise N(0,σ^2): σ=3.4 MU)

The **pure tones** had a centre frequency of 1000 Hz. The duration of the tones was set to 800 ms and they included 125 ms cosine ramps. The **broad-band noises** had a flat frequency response between 100 and 10000 Hz. The duration of the noises was set to 500 ms and they included 50 ms cosine ramps.

Reference data

The reference data for increment-discrimination thresholds obtained using the PEMO model can be found in (Jepsen et al., 2008, their Fig. 3) (not shown in this appendix).

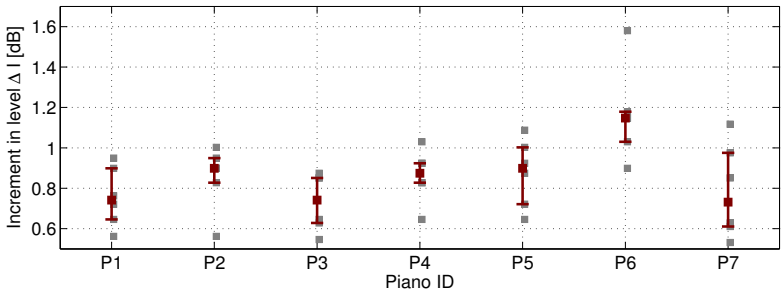


Figure D.4: Results of the increment-detection task simulated using the seven Viennese piano sounds. The median thresholds together with their IQRs are shown. The grey squares indicate the threshold estimation for each staircase procedure.

Simulation results

A variability of $\sigma = 3.4$ MU for the internal noise was first obtained to reach a discrimination threshold $\Delta L = 0.5$ dB for the pedestal level $L = 60$ dB SPL using pure tones. The results for the intensity-discrimination task were then obtained for all 7 pedestal levels for pure tones and broadband noises. Those results are shown in panels (a) and (b) of Figure D.3. The obtained σ value was used to replicate 5 of the 6 auditory tasks evaluated by Jepsen et al. (2008) using the PEMO model, which are shown later in this appendix.

A lower or higher variability of the internal noise (given by its standard deviation σ) would lead to lower (more sensitivity of the model) or higher JNDs in level (less sensitivity of the model), respectively. This is particularly important for the evaluation of deterministic stimuli (e.g., pure tones) or when the same sound excerpt is evaluated repeatedly (e.g., our set of piano sounds). For instance, to increase the JND from $\Delta L = 0.5$ dB (as just reported) to $\Delta L = 1$ dB[1] for the 60-dB pure tone, a standard deviation of $\sigma = 6.7$ MU is required.

D.3.3 Evaluation of piano sounds

(Obtained standard deviation of the internal noise N(0,σ^2): σ=10.1 MU)

The same C#$_5$-note recordings played on the Viennese pianos described in Chapter 3 and 4 were used (see Table 3.1 in Chapter 3). The pedestal (L_{eq}) level of the pianos was not adjusted. The L_{eq} values of the pianos range from 55.4 to 67.2 dB.

Reference data

Due to the high sensitivity of the PEMO model (low ΔL value) when evaluating the intensity-discrimination task using piano sounds and the

[1] A 1-dB criterion was used to calibrate the low-pass modulation model (Dau et al., 1996a, 1996b) and the first versions of the PEMO model (Dau et al., 1997a, 1997b).

obtained σ of 3.4 MU, we decided to decrease the sensitivity of the model to obtain a target discrimination ΔL of 1 dB. We did not collect data to confirm the appropriateness of this criterion. Nevertheless, due to the complex spectro-temporal characteristics of the piano, it is possible that not only another target JND is needed but also a different auditory task. Another auditory task that could be used for setting a limit to the PEMO model is a modulation-increment detection (see, e.g., Ewert & Dau, 2004).

Simulation

The results obtained using an internal (Gaussian) noise with mean $\mu = 0$ and standard deviation $\sigma = 10.1$ MU are shown in Figure D.4. An average discrimination ΔL =0.86 dB across pianos was obtained. The (median) thresholds per piano ranged from 0.73 dB (P7) to 1.15 dB (P6).

D.4 Reproduction of existing simulation data

(Using the internal noise N(0,σ^2) with σ=3.4 MU)

D.4.1 Tone-in-noise experiment

The target sounds were pure tones with a centre frequency of 2000 Hz and durations of 5, 15, 20, 35, 50, 100 and 200 ms. The sounds had 2.5 ms raised-cosine ramps. The sounds were temporally centred in the masker. The masker was a running Gaussian noise limited to the frequency range between 20 and 5000 Hz. The masker had a duration of 500 ms ramped up and down with 10 ms cosine ramps.

Adjustable parameter: level L of target (tone) sounds.
Starting value: $L = 75$ dB.
Number of reversals: 12 (6 reversals in the measurement phase).
Suprathreshold level: $L_{\text{supra}} = 85$ dB

Reference data

The reference data for this task using the PEMO model can be found in (Jepsen et al., 2008, their Fig. 4) (not shown in this appendix).

Simulation

The results for the tone-in-noise task using the PEMO model are shown in panel (a) of Figure D.5.

D.4.2 Forward masking

The masker was a Gaussian noise with frequencies between 20 and 8000 Hz with a duration of 200 ms including 2 ms raised-cosine ramps. The

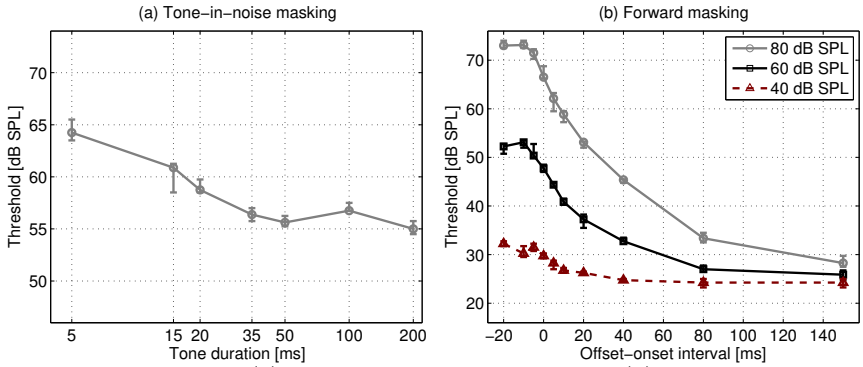


Figure D.5: Results of the (a) tone-in-noise experiment, and (b) the forward-masking experiment at three masking levels.

masker level was set to either 40, 60 or 80 dB. The signal was a 4000-Hz pure tone with a duration of 10 ms having a Hanning window applied over its entire duration. The tone had a temporal separation between the masker offset and the signal onset of either -20, -10, -5, 0, 5, 10, 20, 40, 80 or 150 ms. The separations between -20 and -5 ms correspond to simultaneous masking conditions, while from 0 to 150 ms correspond to forward masking conditions.

Adjustable parameter: level L of target (tone) signals.
Starting value: $L_{\text{supra}} = L_{\text{masker}} + 10$ dB.
Number of reversals: 12 (6 reversals in the measurement phase).
Suprathreshold level: $L_{\text{supra}} = L_{\text{masker}} + 10$ dB.

Reference data

The reference data for this task using the PEMO model can be found in (Jepsen et al., 2008, their Fig. 6) (not shown in this appendix).

Simulation

The results for the tone-in-noise task using the PEMO model are shown in panel (b) of Figure D.5.

D.4.3 Forward masking: Growth-of-masking

This experiment was set-up as a forward masking task with pure tones. Two conditions were tested: on-frequency listening (tone and masker were in the same band, in this case both tones had a frequency of 4000 Hz) and off-frequency listening (the tone had a frequency of 4000 Hz, the masker had a frequency of 2000 Hz). The detection threshold for the tone level was determined at different masker levels: 30 to 80 dB in steps of 10 dB for the on-frequency listening condition and 60, 70, 80, 85 dB

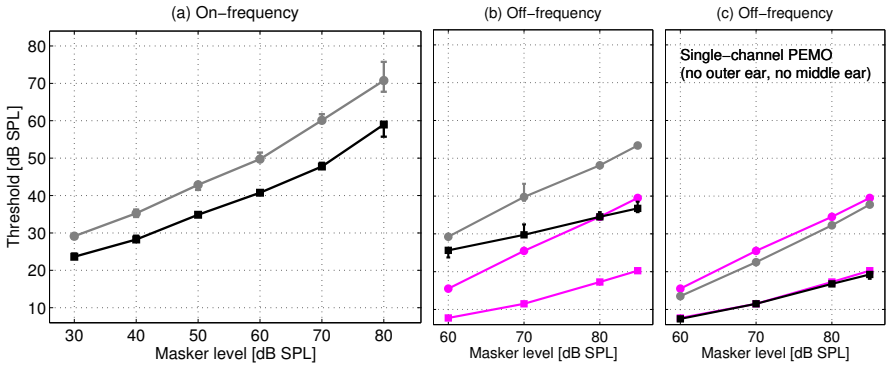


Figure D.6: Results of the growth-of-masking curves in a forward masking experiment using (a) an on-frequency masker, and (b,c) an off-frequency masker. Due to the systematic deviation of the thresholds in panel (b) with respect to the literature (magenta markers), simulations obtained with the PEMO model using only one audio frequency channel (f_c = 4000 Hz) and no outer and middle ear filters are shown in panel (c). The grey curves indicate the detection thresholds for the target sounds staring right after the offset of corresponding masker. The black curves indicate the detection thresholds for the target sounds starting 30 ms after the offset of the corresponding masker.

for the off-frequency listening condition. The signal onset occurred either 0 ms or 30 ms after the masker offset.

Adjustable parameter: level L of target signals.
Starting value: $L_{\text{supra}} = L_{\text{masker}} + 10$ dB.
Number of reversals: 12 (6 reversals in the measurement phase).
Suprathreshold level: $L_{\text{supra}} = L_{\text{masker}} + 10$ dB

Reference data

The reference data for this task using the PEMO model can be found in (Jepsen et al., 2008, their Fig. 7). Detection thresholds for (only) off-frequency maskers from the literature are indicated by the magenta curves in panels (b) and (c).

Simulation

The results for the growth-of-masking experiment in a forward masking task using the PEMO model are shown in Figure D.6. The detection thresholds obtained using on-frequency and off-frequency maskers are shown in panels (a) and (b), respectively. The thresholds shown in panel (b) are on average 13.8 and 17.6 dB above the thresholds from the literature (magenta markers) for the signal onsets 0 and 30 ms after the masker offset, respectively. The simulations were re-run using the PEMO model in a single-channel configuration and bypassing the outer and middle ear filtering. This is the configuration of the model reported

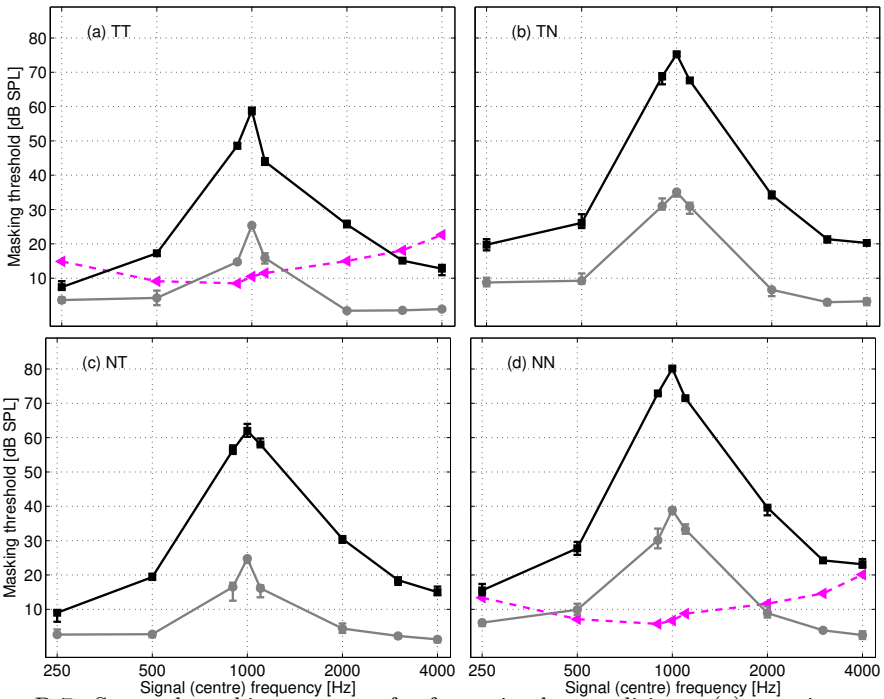


Figure D.7: Spectral masking patterns for four stimulus conditions: (a) tone-in-tone TT, (b) tone-in-noise TN, (c) noise-in-tone NT, and (d) noise-in-noise NN. In panels (a) and (b) simulated absolute thresholds for tones T and noises N are indicated by the magenta dotted lines, respectively.

for the PEMO model data (Jepsen et al., 2008). The results of this simulation are shown in panel (c) of Figure D.6.

D.4.4 Simultaneous masking patterns

The target and reference signals were either a tone or an 80-Hz wide running Gaussian noise with a duration of 220 ms and 10 ms raised-cosine ramps. The signals had a centre frequency of 250, 500, 900, 1000, 1100, 2000, 3000, and 4000 Hz. The masker was always centred at 1000 Hz and it had a level of 45 or 85 dB. There were four possible target-reference signal combinations: (1) tone signal and tone masker (TT), (2) tone signal and noise masker (TN), (3) noise signal and tone masker (NT), and (4) noise signal and noise masker (NN). In the TT condition the masker had a 90° phase shift. For the other conditions random phases were used.

Adjustable parameter: level L of target signals.
Starting value: $L = 75$ dB.
Number of reversals: 12 (6 reversals in the measurement phase).
Suprathreshold level: $L_{\text{supra}} = L_{\text{masker}} + 10$ dB

Reference data

The reference data for this task using the PEMO model can be found in (Jepsen et al., 2008, their Fig. 5) (not shown in this appendix) and they are reported as masked thresholds which are obtained as the detection thresholds in dB SPL referenced to the absolute threshold of hearing for the target signals.

Simulation

The results for the simultaneous-masking experiment using the PEMO model are shown in Figure D.7. The results are shown as masked thresholds in dB to allow a direct comparison with values from the literature. First the absolute thresholds for the target signals (tones T or noise N) centred at the test frequencies were obtained, which are indicated by the magenta dotted lines in panels (a) and (d) for tone T and noise N targets. The masked thresholds were obtained by subtracting those absolute thresholds from the simulated detection thresholds for the four target-reference signal combinations. The resulting curves are shown in panels (a-d) for tone-in-tone, tone-in-noise, noise-in-tone, and noise-in-noise conditions, respectively.

E Auditory modelling: Other approaches to assess the memory template

This appendix contains a description of different template approaches that were evaluated during the development of a central processor for the PEMO model in the context of the perceptual similarity task described in Chapter 3 and simulated in Chapter 4. The finally adopted template approach is described in Chapter 4. This appendix is devoted to the description of those template approaches that did not lead to a satisfactory explanation of the experimental results of Chapter 3. We believe, however, that it is worthwhile to report these approaches indicating the reasons we had to leave them aside.

We start by providing some theoretical background behind the idea of memory templates in the context of an optimal detector (see Green & Swets, 1966, their Chapters 6 and 7). This is followed by a description of the criteria we used to choose possible template approaches. We finally describe two of these alternative approaches and report the argument that lead us to discard them.

E.1 Theory for the derivation of a memory template

In a 3-AFC task approached using an artificial listener (in this dissertation the PEMO model), the three trial intervals can be compared with an "expected signal" or template T_p. If the representations of each interval are labelled as R_x with $x = 1, 2, 3$, then the template is derived from the representation that is related to the target sound ($R_{x,t}$) at a condition that is easy to discriminate, i.e., at a condition that is above threshold (suprathreshold condition). In a detection-in-noise experiment, such a condition is given when the background noise is low (i.e., high SNR) which, by convention, is indicated as $R_{x,t}(MT)$. In the course of a simulated experiment the artificial listener chooses as target interval $R_{x,t}$

(that may be correct or not) the interval that has the highest similarity with T_p. One mathematical way to express this idea is to assess the CCV value between R_x and T_p. The expression to assess the CCV value in continuous and discrete time domains is given by:

$$\text{CCV}_x = \int_0^T R_x(t) \cdot T_p dt \approx \sum_{n=1}^{N} R_x[n] \cdot T_p[n] \Delta t \tag{E.1}$$

In a simplified form:

$$\text{CCV}_x = \frac{1}{f_s} \sum_{n=1}^{N} R_x[n] \cdot T_p[n] \tag{E.2}$$

where f_s is the sampling frequency of the internal representations R_x and T_p. The representations R_x and T_p are N-samples long. This operation can be interpreted as a "template weighting" and is referred in the literature to as template-matching. The assessment of CCV values can in fact be performed along more dimensions of R_x and T_p as long as the samples $R_x[n]$ and $T_p[n]$ in the product of Equation E.2 belong to the same dimension. In general, the internal representations using the PEMO model have three dimensions: time, audio frequency, and modulation frequency.

In Equation E.2, the template T_p sums up (or subtracts) the parts of the representation R_x that have the same (or a different) sign, emphasising them (or de-emphasising them) by an amount defined by the sample-by-sample amplitudes of T_p. It is important to note that, in order to introduce an adequate weighting to the representation R_x, the template T_p should have unit energy.

The template approaches described in this appendix consider different ways to use Equation E.2: (1) by using R_x (as shown in the equation) or ΔR_x (i.e., subtracting the noise-alone representation), and (2) by using signed or unsigned samples in the equation.

E.1.1 Template weighting: Normalisation of the template

One property that has to be satisfied by the derived template T_p is to have unit energy (Dau et al., 1996a):

$$E = \int_0^T T_p^2(t) dt \approx \frac{1}{f_s} \sum_{n=1}^{N} T_p^2[n] = 1 \tag{E.3}$$

where the left hand expression assumes a template T_p in the continuous time domain and the right hand expression in the discrete time domain. The constants T and N represent the duration of the template in seconds and in samples, respectively. The discrete representation has a time resolution $\Delta t = 1/f_s$ [s], with f_s being the sampling frequency of the model representation in Hz.

To derive a template meeting the condition imposed by Equation E.3, a scaled representation of the target interval $R_{x,t}(MT)$ at a suprathreshold SNR can be obtained. In this way, the template has the form $T_p = c \cdot R_{x,t}(MT)$, and the constant c can be obtained as:

$$c = \sqrt{\frac{f_s}{\sum_{n=1}^{N} R_{x,t}^2[n]}} \tag{E.4}$$

E.2 Criteria to be met

E.2.1 Template in a similarity task

The derivation of the template T_p in a similarity task where two (piano) sounds are compared, as described in Section 3.2.3, must be somehow related to: (a) the two test sounds, the target and "reference" pianos, and; (b) two or more realisations of a noise that can efficiently mask the properties of both piano sounds. To account for the latter aspect, noise is always added in every piano presentation (in this thesis they are ICRA noises). For the first aspect, the internal representations of the target piano R_t needs to be used but the representation of the reference piano R_r might also be needed, because the discrimination between pianos depends on how different they are from each other.

Finally, the internal representation of the noises alone R_N might also be used in the template derivation. Despite the fact that in the instrument-in-noise test, noise alone conditions are never presented, the listener might be able to evaluate the similarity among intervals based on how prominent the reference and target piano sounds are with respect to the (ICRA) noises.

E.2.2 Maximisation of the correlation between the template and the internal representations

It is relevant that the template T_p is maximally correlated with each of the intervals R_x because it may be expected that human listeners try to maximise the match between the expected signal (that we assumed to

be "learned") and each of the intervals that are heard. To maximise the CCV values, different time alignments of the involved internal representations should be evaluated during either: (a) the template derivation, or (b) the correlation between the template and target and reference intervals. The relevance of this aspect relies on the fact that the template T_p and the representations R_x are digitised signals, which are sensitive to any eventual misalignment among them. This is in contrast to the rationale of a memory template, where the awareness of the artificial listener about the target signal should be independent of the specific moment, i.e., the specific time alignment, when the test sounds are heard.

E.2.3 Compatibility of the template approach

The template approach should be compatible with the auditory tasks described in Appendix D. This is motivated by the fact that a detection-in-noise task could also be seen as a similarity task, where a comparison is made between the three intervals (signal-plus-noise and two noise-alone intervals) and the template. The comparison is based on CCV values and the artificial listener chooses as the interval containing the target sound the interval that produces the highest CCV value, i.e., the "most similar" interval with respect to the template.

E.2.4 Adjustment of the sensitivity of the artificial listener

The use of different template approaches may introduce changes in the sensitivity of the artificial listener. To compensate for eventual changes in the sensitivity of the artificial listener (i.e., the PEMO model), the variability σ of the internal noise is checked and adjusted (if needed) by re-running the increment-detection experiment described in Section D.3.3.

E.3 Simulation procedure

All simulations were run using the AFC toolbox for MATLAB (Ewert, 2013). In this toolbox an artificial listener was enabled to conduct the listening experiments. The artificial listener processed the (whole-duration) sounds using the auditory PEMO model with the set of parameters listed in Chapter 4 using two overshoot limitation factors ($\mathsf{limit}= 10$ and $\mathsf{limit}= 5$) in the adaptation loop stage. The experiments were all implemented as 3-AFC tasks using a two-down one-up tracking rule.

For each template approach, the experiments were always run in the following order: (1) Increment-detection using C#$_5$ (anechoic) piano

sound (see Section D.3.3), (2) instrument-in-noise experiment, and (3) forward-masking experiments (see Section D.4.2). The first experiment was run to assess the amount of variance σ needed for the internal noise of the central processor, the second experiment was run to evaluate the artificial listener's performance in our instrument-in-noise task. The third experiment was run to evaluate the compatibility of the adopted approach with the estimation of forward-masking thresholds. The forward-masking experiment was chosen with the motivation to replicate the threshold estimation of one of the detection-in-noise tasks reported in Appendix D.

E.3.1 Stimuli

The same C#$_5$-note (anechoic) recordings played on the Viennese pianos described in Chapter 3 and 4 were used for the simulation of the instrument-in-noise experiment. Only a subset of 9 piano pairs (of the 21 possible combinations) were used. The selected 9 piano pairs are well distributed along the experimentally-obtained scale of similarity and they were also used in the exploratory simulations presented in Chapter 4. The selected piano pairs were: pair 12, 15, 16, 23, 26, 27, 37, 45, and 47.

E.4 Approach 1: Piano-plus-noise templates

Description

In this approach one template is used. The template T_p is derived from the representation of the interval that contains the target piano sound ("target piano-plus-noise" interval). The approach is very similar to the derivation of templates that has been adopted so far in the literature (see, e.g., Dau et al., 1996b, 1997a; Jepsen et al., 2008). The target piano-plus-noise interval (presented once) is treated as the signal-plus-noise interval of a detection-in-noise experiment. Correspondingly the reference piano-plus-noise intervals (presented twice) are treated as the noise-alone intervals of the detection task.

In this approach the CCV between the template T_p and the piano-plus-noise sounds for the intervals $x = 1, 2$, and 3 were obtained using two variants:

$$\mathrm{CCV}_x = \frac{1}{f_s} \sum_{n=1}^{N} R_x[n] \cdot T_p[n] \tag{E.5}$$

and

$$\mathrm{CCV}_x = \frac{1}{f_s} \sum_{n=1}^{N} \Delta R_x[n] \cdot T_p[n] \tag{E.6}$$

Criterion of the artificial listener

If interval $x = 1$ of the 3-AFC trial contains the target piano, then the artificial listener makes a correct decision if:

$$\max \left\{ \widehat{\mathrm{CCV}}_{x,t} \right\} = \widehat{\mathrm{CCV}}_{1,t} \tag{E.7}$$

The hat symbol indicates that internal (Gaussian) noise $\mathrm{N}(0,\sigma^2)$ is added to the CCV_x values before the artificial listener makes a decision.

Why not use this approach

The simulated thresholds $\mathrm{thres}_{\mathrm{sim}}$ ranged from -7.0 to 2.5 dB for **variant 1** (Equation E.5) and from -3.5 to 5.5 dB for **variant 2** (Equation E.6). These reduced ranges of threshold values contrast with the range of experimental thresholds from $\mathrm{thres}_{\mathrm{exp,min}} = -1.75$ dB and $\mathrm{thres}_{\mathrm{exp,max}} = 20.75$ dB that is reported in Chapter 3. Due to this large discrepancy and because in this approach the template derivation used only the target piano representation R_t, we decided to add explicit information of the reference piano R_r in **Approach 2**.

E.5 Approach 2: Difference representation

Description

In this approach, the representation of the reference piano R_r is subtracted from the representation of the target piano R_t. The difference representation $\Delta R = \|R_t - R_r\|$ is further analysed. The difference representation is used now as a quantitative distance measure between representations. Another study where an unsigned difference between internal representations has also been used is given by Agus et al. (2012). The expression to assess the CCV values using the difference representations has to be adjusted, because the artificial listener does not know which interval contains the target and which of the other two intervals contains the reference signals. The expression to obtain the CCV value can be written as follows:

$$\mathrm{CCV}_{xy} = \frac{1}{f_s} \sum_{n=1}^{N} \|\Delta R_{xy}[n]\| \cdot \|T_p[n]\| \tag{E.8}$$

the subindexes x and y indicate that the representation R_y from the interval y is subtracted from the representation R_x from the interval x ($\Delta R_{xy} = R_x - R_y$).

Criterion of the artificial listener

Three CCV values are obtained using Equation E.8 using the internal representations of intervals $x = 1, 2$, and 3 namely CCV_{12}, CCV_{13}, and CCV_{23}. If the template T_p has also been derived from a difference representation between target and reference sounds, and we assume that interval 1 contains the target sound, then ΔR_{12} and ΔR_{13} should produce a higher CCV than ΔR_{23}. This is because ΔR_{23} does not account for the representation of the target sound. One way to translate this into a discriminability outcome is to look for the lowest CCV value (in the example CCV_{23}). The artificial listener then chooses the "other" interval as the target interval (in the example interval $x = 1$).

Why not use this approach

The simulated thresholds $\text{thres}_{\text{sim}}$ had a similar range of values compared with those reported for the two variants of Approach 1, from -8.0 to 3.75 dB. We faced, however, an additional problem for generating difference representations ΔR_{xy} namely to find out a systematic way of ensuring maximum (and reliable) CCV values between ΔR_{xy} and the template. Different "types of difference representations" need to be generated during the simulation of the similarity task, namely for (1) deriving the template, (2) deriving the difference between target and reference representations (R_{12} and R_{13}), and (3) deriving the difference between reference representations R_{23}. For each of those cases a different alignment of the internal representations can increase or decrease the obtained CCV values.

In order to try another approach where both the target and reference piano representations can be used by the artificial listener but, at the same time, reducing the dependency of the model judgements on finding an appropriate alignment criterion, we decided to adopt a criterion similar to that of **Approach 1** but using two templates: T_p as in **Approach 1** (labelled as $T_{p,t}$) and another template derived in a similar way from the reference piano sound (labelled as $T_{p,r}$). Such a template was adopted and further investigated in Chapter 4.

Acknowledgements

This dissertation is the end result of a four-years path along which I was lucky to always be surrounded by good people. Without all their support this project would have not been the same. For this reason I dedicate the following lines to the many people who accompanied me along this path.

Firstly I would like to thank my supervisor. Armin, thank you for all the time you spent on our long weekly discussions, the enthusiasm that you always showed and the flexibility to meet in different places at different times not only including our offices but also our homes, cafés and restaurants. I believe this dissertation reflects many of the things that I have learned from you in these four years.

I am also very grateful to my co-supervisor Antoine. I am very honoured to have worked with you. I loved working with your piano recordings and to get familiar with some historical and technical facts around the piano construction. You always made time for me even during your busy period in Vienna. I would also like to thank all other colleagues and friends from our European project BATWOMAN especially Winfried, Eckard, Sebastià and Malte.

Much of my gratitude goes to my colleagues and friends at the HTI group, for all the nice moments within and outside IPO. In particular I would like to thank Kong, Chao, Mark, Heleen, Giacomo, Kevin, Toros, Anne, Indre, Hanne, Maaike, Laura, Els, Elçin, Milou, Patty, Samantha, Minha, Caixia, Sofia, Alain, Peder, Anne M. and also to Leon, Sheng, Mieke, Mariska, Peter, Frank and Renske for the many coffees together, walks around, Cookie Wednesdays, and other spontaneous get-togethers. I am also thankful for the open-door approach in our group and for the help that I repeatedly got from Ellen, Dik, Daniël, Peter, Raymond, Martin and Aart. My office mates also deserve a special mention: Ryan, Huihui, Nemanja, Rebecca, Sima and Margot. Thank you for the good time at IPO 0.21. I would also like to thank the students I had the opportunity to supervise in particular Rodrigo, Kevin and Glen, who collected part of the data used in this thesis.

During my Ph.D. I regularly travelled between Antwerp and Eindhoven. Despite the long travelling times, I can count two positive consequences. The first one is that I was able to meet two very good friends. Ake en Casimir, it is nice to realise that although you do not need to come to Eindhoven anymore, we have managed to keep the contact with each other. The second consequence was that I got to know better some Eindhoven-friends every time I could benefit from their hospitality. Those stays in Eindhoven were most of the times complemented with either dinners, drinks, Champions League matches or a combination of them. Thank you Kong, Toros, Peder, Edgar and Llaima.

Federico and Michael, I am also grateful to you and our long lasting friendship. I do not only enjoy every time we meet, but I also get inspired when we discuss about our ongoing research or any other "geek" stuff.

I am eternally grateful to my family in Chile. My parents Luis and Alicia have always stayed close to me despite the long distance that separates us. Papito y mamita, muchas gracias por el apoyo incondicional que siempre me han dado, gracias por todos los sacrificios que ustedes hicieron para que yo llegue a donde estoy, ustedes siempre han sido y seguirán siendo mi ejemplo a seguir. Thank you Carolina and Roberto for shortening the long distance with our spontaneous chats. Caro, gracias por siempre hacerme partícipe de tu vida, incluso ahora que estamos tan lejos. Me siento orgulloso de que ahora estés haciendo lo que realmente te gusta y de la vida que estás formando con Manuel. Rober, los últimos años te han tocado duro, pero pese a eso siempre has estado cuando lo he necesitado. Me alegro de verte junto a Mónica y de tus ahora tres preciosos hijos León, Matías y Rafael.

I am also grateful to my in-laws, mijn schoonouders Gilbert en Gert, mijn schoonbroers en hun wederhelft Jeroen en Karen, Stijn en Iris, mijn schoonnichtjes (echt schoon) Jade, Linde en Evelyn en mijn neef (ook heel schoon) Lenn. Jullie hebben me met open armen ontvangen en aanvaard zonder daar iets voor terug te vragen. Bedankt voor jullie steun.

Finally I would like to thank my wife Frauke. We have done so much in the last years. You and our two beautiful children Paula and Alexis are for me the biggest source of inspiration, motivating me to become every day better. Pingüinita, gracias por estar a mi lado. Tu apoyo, comprensión, empatía, amistad y cercanía me hacen un mejor hombre, padre, persona y también investigador. Por esto, sin darte cuenta, tú eres quien más ha contribuido al trabajo que presento en esta tesis.

Curriculum Vitae

Alejandro Osses Vecchi was born on 20 May 1985 in Santiago, Chile. He obtained the professional degree of *civil engineer in sound and acoustics* in 2010 at the Technological University of Chile INACAP (former Vicente Pérez Rosales). He then worked in the area of environmental acoustics as project manager and project engineer at the consultancy companies Acustical (2010-2012) and Control Acústico (2011-2012), respectively. In 2012 he moved to Belgium, where he completed a predoctoral programme in *biomedical sciences* at the University of Leuven. The topic of his research project was concerned with real-time audio signal processing for cochlear implants. Since May 2014 Alejandro is part of the Human-Technology Interaction group at the Eindhoven University of Technology where he started a Ph.D. project under the supervision of prof. Armin Kohlrausch. The project titled "Prediction of perceptual similarity based on time-domain models of auditory perception" was performed within the Initial Training Network BATWOMAN in the framework of a Marie Skłodowska-Curie Action, with the main goal to stimulate interdisciplinary research among three disciplines –musical acoustics, room acoustics, and automotive applications– with perception as a central focus.

Publications

Peer reviewed papers

A. Osses, and A. Kohlrausch (2018, submitted). "Auditory modelling of the perceptual similarity between piano sounds." *Acta Acust. united Ac.*

A. Chaigne, A. Osses, and A. Kohlrausch (2018, submitted). "Similarity of piano tones: a psychoacoustical and sound analysis study." *Applied Acoustics*.

A. Osses, A. Kohlrausch, W. Lachenmayr, and E. Mommertz (2017). "Predicting the perceived reverberation in different room acoustic environments using a binaural auditory model." *J. Acoust. Soc. Am.* 141(4), pp. EL381-EL387. doi:10.1121/1.4979853.

Papers in preparation

A. Osses, A. Kohlrausch, and A. Chaigne. "Perceptual similarity between piano notes: Experimental data for reverberant and non-reverberant sounds."

A. Osses, and A. Kohlrausch. "Perceptual similarity between piano notes: Simulations with a template-based perception model."

Non-peer reviewed papers

A. Osses, A. Chaigne, and A. Kohlrausch (2017). "Meten van klankverschillen in klassieke piano's" (Measurement of sound differences in classic pianos, in Dutch). *Nederlands Tijdschrift voor Natuurkunde* 83 (7), pp. 246-249.

Osses, A., Chaigne, A. and Kohlrausch, A. (2016). Assessing the acoustic similarity of different pianos using an instrument-in-noise test. International Symposium on Musical and Room Acoustics, pp. 1-10. La Plata, Argentina. (Link to download)

Osses, A., García, R. and Kohlrausch, A. (2016). "Modelling the sensation of fluctuation strength. *Proc. Mtgs. Acoust.* 28 (50005), pp. 1-8. doi:10.1121/2.0000410.

A. Osses, C. Kim, and A. Kohlrausch (2015). "Perceptual evaluation of differences between original and synthesised musical instrument sounds: the role of room acoustics." Proceedings of EuroNoise. Ed. by C. Glorieux. Maastricht, the Netherlands, pp. 2561-2566. (Link to download)

Colophon

This thesis was typeset using LaTeX. The cover of this dissertation was designed by Carolina Osses Vecchi. This dissertation was printed by: ProefschriftMaken || www.proefschriftmaken.nl