Perceptual similarity between piano notes

Citation for published version (APA):

DOI:
10.1121/1.5121311

Document status and date:
Published: 01/08/2019

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

Please check the document version of this publication:

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

Link to publication

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

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

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

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Perceptual similarity between piano notes: Experimental method applicable to reverberant and non-reverberant sounds\textsuperscript{a)}

Alejandro Osses Vecchi\textsuperscript{b)} and Armin Kohlrausch

\textit{Human-Technology Interaction group, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, 5600MB Eindhoven, The Netherlands}

Antoine Chaigne

\textit{Institute of Music Acoustics, University of Music and Performing Arts, Vienna, Austria}

(Received 31 January 2019; revised 16 July 2019; accepted 22 July 2019; published online 8 August 2019)

In this paper an experimental method to quantify perceptual differences between acoustic stimuli is presented. The experiments are implemented as a signal-in-noise task, where two sounds are to be discriminated. By adjusting the signal-to-noise ratio (SNR) the difficulty of the sound discrimination is manipulated. If two sounds are very similar already, a low level of added noise (high SNR) makes the discrimination task difficult. For more dissimilar sounds, a higher amount of noise (lower SNR) is needed to affect discriminability. In other words, a strong correlation between SNR and similarity is expected. The experimental noises are generated to have similar spectro-temporal properties to those of the test stimuli. As a study case, the suggested method was used to evaluate recordings of one note played on seven Viennese pianos using (1) non-reverberant sounds (as recorded) and (2) reverberant sounds, where reverberation was added by means of digital convolution. The experimental results of the suggested method were compared with a similarity experiment using the method of triadic comparisons. The results of both methods were significantly correlated with each other. \textcopyright 2019 Acoustical Society of America. https://doi.org/10.1121/1.5121311

I. INTRODUCTION

Perceptual similarity between stimuli is a problem approached in several disciplines and is normally assessed experimentally. Popular experimental tasks used to compare sounds are the methods of triadic comparisons (Fritz et al., 2010; Levelt et al., 1966; Novello et al., 2011), pairwise comparisons (Grey, 1977; Grey and Gordon, 1978; Raake et al., 2014; Tahvanainen et al., 2015), and free verbalization rating and categorization (Dubois, 2000; Guastavino and 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 (Osses, 2018, Chap. 1). For the methods of triadic\textsuperscript{1} 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 such as multidimensional scaling (MDS; Kruskal, 1964b; Shepard, 1962) and the use of the Bradley-Terry-Luce (BTL) model (Bradley, 1953; Wickelmaier and Schmid, 2004) are examples of algorithms that allow such a comparison. Our interest was, however, not only in knowing which sounds are more or less similar but also in obtaining a way to manipulate their perceptual distances. In this paper we show a way to reach that objective by conducting a listening test to discriminate two sounds using a three-alternative forced-choice (3-AFC) experiment in noise, where the noise allows to change the similarity of the sounds being tested. In Sec. II, the discrimination test–that in the context of this paper we named instrument-in-noise test–is explained, providing a detailed description of the noise generation. As study case, a comparison of one note (C\#5) recorded from seven Viennese pianos built in the 19th century is given. The experimental procedure used to test the piano sounds is explained in Sec. III, where also a description of the reference method of triadic comparisons is given. In Secs. IV and V, the experimental results and data analysis are presented and, finally, a summary is presented in the Conclusion, where the method is put into a broader context.

II. DESCRIPTION OF THE METHOD

A method to quantify the perceptual differences between sounds, expressed in signal-to-noise ratio (SNR), is presented in this section. The sounds are compared pairwise, and they are embedded in a background noise. 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 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 indistinguishable. To deliver such results, we hypothesized that the noise needs to

\textsuperscript{a}Portions of this work were presented in “Assessing the acoustic similarity of different pianos using an instrument-in-noise test,” International Symposium on Musical and Room Acoustics, La Plata, Argentina, 2016.

\textsuperscript{b}Currently at: Hearing Technology @ WAVES, Dept. of Information Technology, Ghent University, 9052 Zwijnaarde, Belgium. Electronic mail: ale.a.osses@gmail.com
have similar spectro-temporal properties as 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 more accurate 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 piano sounds.

A. Modified ICRA noise

The procedure to generate ICRA noises introducing a musical-instrument weighting is illustrated in Figs. 1 and 2 for two variants of the algorithm, versions A and B. One fundamental change in these modified ICRA algorithms compared to the original description by Dreschler et al. (2001) is the use of a Gammatone filter bank instead of the original band-split filter with crossover frequencies at 800 and 2400 Hz. For speech signals, the resulting frequency bands were chosen to manipulate three relevant frequency regions related to the fundamental frequency and second formant of voiced segments, and the range of unvoiced fricatives. The use of a Gammatone filter bank provides more freedom to follow the spectral properties of the input sounds.

Our two variants of the ICRA noise algorithm introduce a similar temporal weighting but they differ slightly in the spectral weighting. Version B follows more accurately the spectral properties of the sounds. The spectral weighting in version A presents an increasing spectral tilt that introduces a gradual band weighting of up to 10 dB at the highest auditory filter with respect to the \( f_0 \)-centered band. We only became aware of this effect after running the experiments using noise A. We compensated the spectral tilt in the algorithm version B. Although we do not include an in-depth analysis between experimental results using versions A and B, some reflection about the influence of the spectral tilt is given in Sec. VC, and this can be further investigated using computational modelling (Osses, 2018).

1. Version A

a. Stage A.1. Band-split filter. The input signal is fed into a Gammatone filter bank that consists of 31 bands with center frequencies between 87 Hz (3 ERB\(_N\)) 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 AMT toolbox for MATLAB was used for this purpose (Søndergaard and Majdak, 2013). The filter design and processing introduced in this stage are equivalent to the “frequency analysis” stage described by Holmann (2002).

b. Stage A.2. Sign randomization. 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). The resulting waveforms have a flat spectrum while keeping the same temporal envelope characteristics and same band level.

c. Stage A.3. Re-filtering per band-split filter. Each \( i \)th signal is fed into the \( i \)th band of the gammatone filter bank. The index \( i \) represents each of the 31 bands.

d. Stage A.4. Add signals together. The 31 filtered signals are added together.

e. Stage A.5. Phase randomization. The phase of the signal is randomized following a uniform distribution between 0 and \( 2\pi \). This is done in the frequency domain using a 512-point fast Fourier transform (FFT; 87.5% overlap) by overlap/adding segments after an inverse FFT (IFFT) is applied. The resulting signal is adjusted to have the same total root-mean-square (RMS) level as the input to the band-split filter stage.

f. Stage A.6. Low-pass filter. The signal is filtered using an eighth-order Butterworth filter with a cutoff frequency at the upper limit of the highest critical band (\( f_{\text{cutoff}} = 8200 \) Hz).
This filter is introduced to reduce undesired high frequencies as a consequence of the phase randomization.

2. Version B

a. Stage B.1. Band-split filter. The input signal is fed into a gammatone filter bank that consists of 30 bands with center frequencies between 100 Hz (3.4 ERB) and 7330 Hz (32.4 ERB), spaced at 1 ERB. This number of bands was obtained by using \( f_0 = 554 \text{ Hz} \) (11.4 ERB) as base frequency. This corresponds to the same all-pole implementation as in version A.

b. Stage B.2. Sign randomization. The sign of each sample of the 30 filtered signals is either reversed or kept unaltered with a probability of 50% (as in version A).

c. Stage B.3. Re-filtering per band-split filter and band-level adjustment. The \( i \)th signal is fed into the \( i \)th band of the gammatone filter bank. 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 levels back to the values as before this stage.

d. Stage B.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 done because the gammatone filter bank is implemented as a set of infinite impulse response (IIR) filters and, therefore, the filter bank has frequency-dependent group delays. The delays being compensated range from 5.6 ms (bands centered at \( f_c = 554 \text{ Hz} \) or below) down to 0.57 ms (\( f_c > 7330 \text{ Hz} \)). Those delays correspond to the time stamp at which each band-pass filter \( (f_c \geq 554 \text{ Hz}) \) has a maximum in its envelope, when an impulse is used as input. For filters with \( f_c < 554 \text{ Hz} \) 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 during the algorithm the signals are filtered (stage B.1) and then re-filtered (stage B.3).

B. 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 (non-reverberant) recordings of note C#5 \( (f_0 = 554 \text{ Hz}) \) from pianos P1 and P3 were chosen (see Table I). First, the sounds were set to have the same duration (1.3 s) with the note onset occurring approximately at the same time stamp (0.1 s). The ICRA noise for each test sound may be generated using either variant of the ICRA-noise algorithm. Version A of the algorithm was used in this example. The resulting ICRA noises have an average (RMS) level that is the same as the level of the corresponding input signals, which can be interpreted as being at an SNR of 0 dB. Pianos P1 and P3, together with one realization of their ICRA noises, are shown in Fig. 3. The sounds are compared pairwise with the task of distinguishing between the two sounds, while being presented in an embedded (ICRA) noise, using a 3-AFC procedure. In the 3-AFC procedure, also known as odd-ball paradigm, one of the two sounds serves as “reference” and is presented in two observation intervals. The other sound serves as “target sound” and is presented in the randomly chosen third interval. The noise in each interval has to account for the spectro-temporal properties of both piano sounds. For this purpose, a “paired noise” with an SNR of 0 dB is generated by combining the ICRA noises of each test sound (mean of their waveforms). This approach is based on the assumption that such a noise is efficient to gradually mask the properties of the test sounds (in the example, P1 or P3 plus the paired noise) when the noise level increases (i.e., the SNR decreases). It is important, however, to use different paired-noise realizations 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 (e.g., von Klitzing and 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, running noises are used. These are noises that are independently generated but being drawn from the same statistical distribution. To generate running ICRA noises, 12 realizations of each paired ICRA noise are generated. Within each 3-AFC trial,

<table>
<thead>
<tr>
<th>Piano information</th>
<th>(1) Non-reverberant condition</th>
<th>(2) Reverberant condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piano/Manufacturer (year)</td>
<td>Level (dB SPL)</td>
<td>Loudness (sone)</td>
</tr>
<tr>
<td>P1/G. Heccher (1805)</td>
<td>77.2/62.8</td>
<td>17.4 (13.7–22.0) / 6.8 (5.2–8.8)</td>
</tr>
<tr>
<td>P2/N. Streicher (1819)</td>
<td>74.9/58.8</td>
<td>17.2 (13.5–21.8) / 5.5 (4.2–7.2)</td>
</tr>
<tr>
<td>P3/G. Graf (1828)</td>
<td>73.7/55.4</td>
<td>17.0 (13.3–21.5) / 5.6 (4.3–7.3)</td>
</tr>
<tr>
<td>P4/J. B. Streicher (1836)</td>
<td>83.7/66.3</td>
<td>18.5 (14.4–23.5) / 7.0 (5.3–9.1)</td>
</tr>
<tr>
<td>P5/J. B. Streicher (1851)</td>
<td>78.0/60.2</td>
<td>17.8 (14.1–22.4) / 6.6 (5.1–8.5)</td>
</tr>
<tr>
<td>P6/J. B. Streicher (1873)</td>
<td>81.7/67.2</td>
<td>17.2 (13.5–21.8) / 7.3 (5.6–9.1)</td>
</tr>
<tr>
<td>P7/J. B. Streicher and Sohn (1873)</td>
<td>81.7/67.2</td>
<td>17.4 (13.7–22.1) / 8.3 (6.3–10.7)</td>
</tr>
</tbody>
</table>

\(^a\) Piano P1 is a contemporary replica of a piano built in 1805.
\(^b\) Pianos P5 and P6 differ in their hammer action: English and Viennese, respectively.
three paired noises are chosen, leading to 220 possible triads of noises (i.e., 12 choose 3 triads). If the selection of noises is 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 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.

C. 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 discrimination threshold is estimated by adjusting the noise level. This corresponds to an adaptive procedure (or staircase method). The participant has to identify which of the three intervals contains the target sound. The adjustable parameter (noise level expressed as SNR in dB) is adjusted following a two-down one-up rule (Levitt, 1971), which tracks the 70.7% discrimination threshold. We chose to wait until 12 reversals are reached before stopping each staircase estimation. The starting point of the paired ICRA noise is set to an SNR of 16 dB. The step size by which the noise is adjusted is set to 4 dB and is reduced to 2 dB (after the second reversal) and 1 dB (after the fourth reversal). After the fourth reversal the runs continue with a fixed step size of 1 dB. These runs correspond to the measuring stage. The median of the reversals during the measuring stage (last eight reversals) are used to estimate the discrimination threshold of each pair of sounds.

The piano sounds that are compared in this paper differ considerably in loudness due to differences in the construction of the instruments, which was affected by fast technological developments during the 19th century. Loudness cues are, however, not the main focus of this paper. To avoid the use of loudness cues during the 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 not use level as discrimination criterion. The intervals lasted either 1.3 or 2 s (see the dataset descriptions below) with an interstimulus interval of 0.2 s. Based on pilot experiments, each staircase was expected to have an average of 45 trials and a duration of about 8 min. With a dataset of seven sounds, the number of pairwise comparisons (with no permutations) is 21 (i.e., 7 choose 2 pairs), meaning that each participant needed about 3 h to evaluate the whole dataset. To reduce this duration, a balanced subset of data was considered (see Sec. III D).

D. Reference procedure: Method of triadic comparisons

The method of triadic comparisons provides a way to obtain similarity judgments 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 (e.g., Burton and Nerlove, 1976; Shepard, 1987), including the assessment of perceptual spaces using sound stimuli (Fritz et al., 2010; Levelt et al., 1966; Novello et al., 2011; van Veen and Houtgast, 1983). We therefore chose this experimental procedure as a reference to validate the instrument-in-noise method.

In the method of triadic comparisons, each trial consists of three sounds (A,B,C). From this triad, it is possible to form three pairs (AB,AC,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 labeled 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
One method to further process the similarity matrix is the MDS algorithm (Kruskal, 1964a,b; Shepard, 1962). MDS is commonly used as a visualization tool of complex data. The similarity matrix is an \( n \times n \) matrix (here, \( n = 7 \)). In the MDS algorithm, the similarity matrix is assigned to a lower-dimensional space \( (n \times q) \) matrix. The Euclidean distances \( d_{ij} \) between elements within this \( q \)-dimensional space correspond to a unidimensional measure of similarity that was used as reference for comparison with the discrimination thresholds of the instrument-in-noise test.

### III. STUDY CASE: SIMILARITY AMONG 19TH-CENTURY VIENNESE PIANOS

#### A. Stimuli

Recordings from seven pianos were compared among each other. The pianos were constructed in Vienna between 1805 and 1873. During this historical period, the piano construction underwent major developments. The most important change 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 into air, increased in thickness to withstand 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 et al., 2019). These changes affected the timbre of the radiated piano sounds. We believe that these seven pianos are a representative sample of the timbre changes of the instrument.

#### 1. Recordings

Recordings of one note (C#5, nominal \( f_0 = 554 \) Hz) from the seven pianos were used. One recording per piano was chosen leading to a total of seven stimuli. The recordings were obtained by A.C. using a ROGA RG-50 1/4 in. microphone connected to LabView at a sampling frequency of 51 200 Hz (Chaigne et al., 2019). The microphone was placed 50 cm in front of the middle-range keyboard and 50 cm above the soundboard. The recordings had durations between 1.4 s (P1) and 3.3 s (P4), and were considered to be nearly anechoic or “non-reverberant.”

#### 2. Preprocessing

First, the recordings were resampled to a rate of 44 100 Hz. For conservation purposes of the recorded pianos, the pitch of their C#5 notes were equal to or lower than the nominal \( f_0 \) of 554 Hz. The mean pitch of the sounds was therefore adjusted to this value. 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 two, a time stretch technique was used to keep the duration of the pitch-adjusted sounds constant. The time stretch was done using the phase vocoder algorithm (Ellis, 2002).

#### 3. Dataset 1: Non-reverberant sounds

The duration of the seven piano sounds 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 model (Glasberg and Moore, 2002) was used. After the adjustment, the individual piano sounds had a maximum level \( L_{\text{max}} \) ranging from 73.7 to 83.7 dB sound pressure level (SPL; Table I).

#### 4. Dataset 2: Reverberant sounds

The seven piano sounds after the preprocessing stage were digitally convolved with a binaural room impulse response (RIR) taken from an existing measurement of a room that had an early decay time \( EDT_{\text{mid}} \) of 3 s. The duration of the convolved sounds was set to 2 s with the note onset occurring at a time stamp of 0.1 s. The sounds were ramped down using a 300-ms linear ramp. This longer-duration ramp was needed to produce a more natural offset of the convolved sounds. As for dataset 1, the loudness of the sounds was adjusted to have a maximum short-term loudness of 18 sone. The resulting sounds had an \( L_{\text{max}} \) ranging from 73.1 to 81.0 dB SPL (Table I).

#### B. Apparatus

The experiments were conducted in a doubled-walled soundproof booth. The stimuli were presented via Sennheiser HD 265 Linear circumaural headphones (Sennheiser, Wedemark, Germany) in a diotic reproduction (identical left and right channels) and stereo reproduction for the experiments with non-reverberant and reverberant sounds, respectively. The participants’ responses were collected on a computer using the APEX software (Francart et al., 2008) for the instrument-in-noise and the APEX Toolbox for MATLAB (De Man and Reiss, 2014) for the triadic comparisons.

#### C. Participants

For each experiment, 20 participants were recruited from the JF Schouten subject database of the Eindhoven University of Technology. They all had self-reported normal hearing, provided their informed consent before starting the experimental session, and were paid for their contribution. This sample size was assessed a priori aiming at testing the hypothesis that the data from the instrument-in-noise test are
highly correlated (correlation of at least 0.60) with the data from the triadic comparisons with a power of 90% and a significance level of 5%. This analysis was done in the software G*Power (Faul et al., 2009), requiring 17 participants to reach the desired effect size. By increasing the number of participants to 20 the observable correlation is reduced to 0.57. Experiments with dataset 1: Participants S01–S20 (8 females, 12 males) were between 19 and 38 years old (median of 25) at the time of testing. Experiments with dataset 2: Participants S01–S20 (3 females, 17 males) were between 19 and 66 years old (median of 24).

D. Experimental sessions

The experimental sessions were organized in two one-hour sessions per participant, including breaks. For the instrument-in-noise test, every participant judged 11 pairs of piano sounds. This means that the whole dataset (21 piano pairs) was tested once every two participants, including one common pair. About three-quarters of the session was needed to evaluate half of the dataset, and the remaining time for judging the triadic comparisons. Participants were encouraged to take breaks if they felt tired or distracted, which may otherwise 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 familiarizing the participants with the set of piano sounds. The session continued with five or six 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 two (or three) staircases were interleaved. This means that the participant did not necessarily judge the same piano pair in consecutive trials. The order of presentation of the 21 possible piano pairs was randomized in a way such that each of them was tested ten times, using either of the sounds within a pair five times as reference and five times as target. This resulted in 220 expected threshold estimations for sessions using non-reverberant sounds (dataset 1), including 5 piano pairs that were tested 12 instead of 10 times. A similar randomization order was followed in the sessions with reverberant piano sounds (dataset 2), but no extra pair evaluations were collected. This means that half of the participants tested 10 pairs and the other half tested 11 pairs with a total of 210 expected threshold estimations. In practice, the experimental sessions using dataset 2 were about 10 min longer compared to the sessions in which dataset 1 was used.

IV. RESULTS

The results of the instrument-in-noise and triadic comparison tests are described first for dataset 1 (non-reverberant sounds) and then, providing less details, for dataset 2 (reverberant sounds). For ease of readability, however, the results for both datasets are presented in the same figure, coded using red and green markers to indicate results using datasets 1 and 2, respectively.

A. Dataset 1: Instrument-in-noise test

The discrimination thresholds of the instrument-in-noise test using the non-reverberant piano sounds are shown in Fig. 4(A). The pooled thresholds [red triangles in Fig. 4(A)] 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 grouped together). The thresholds ranged from 20.75 dB for pair 23 down to −1.75 dB for pair 26, covering an SNR-range of 22.5 dB. The estimations had a large between-subject variability with a length of the interquartile ranges (IQRs) between 3.25 dB (pair 57) and 19.0 dB (pair 23) with a median value of 8 dB. These results were 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 and are being 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 sessions 1 and 2 the same six 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 common pair. About three-quarters of the session was needed to evaluate half of the dataset, and the remaining time for judging the triadic comparisons. Participants were encouraged to take breaks if they felt tired or distracted, which may otherwise 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 familiarizing the participants with the set of piano sounds. The session continued with five or six 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 two (or three) staircases were interleaved. This means that the participant did not necessarily judge the same piano pair in consecutive trials. The order of presentation of the 21 possible piano pairs was randomized in a way such that each of them was tested ten times, using either of the sounds within a pair five times as reference and five times as target. This resulted in 220 expected threshold estimations for sessions using non-reverberant sounds (dataset 1), including 5 piano pairs that were tested 12 instead of 10 times. A similar randomization order was followed in the sessions with reverberant piano sounds (dataset 2), but no extra pair evaluations were collected. This means that half of the participants tested 10 pairs and the other half tested 11 pairs with a total of 210 expected threshold estimations. In practice, the experimental sessions using dataset 2 were about 10 min longer compared to the sessions in which dataset 1 was used.

IV. RESULTS

The results of the instrument-in-noise and triadic comparison tests are described first for dataset 1 (non-reverberant sounds) and then, providing less details, for dataset 2 (reverberant sounds). For ease of readability, however, the results for both datasets are presented in the same figure, coded using red and green markers to indicate results using datasets 1 and 2, respectively.

A. Dataset 1: Instrument-in-noise test

The discrimination thresholds of the instrument-in-noise test using the non-reverberant piano sounds are shown in Fig. 4(A). The pooled thresholds [red triangles in Fig. 4(A)] 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 grouped together). The thresholds ranged from 20.75 dB for pair 23 down to −1.75 dB for pair 26, covering an SNR-range of 22.5 dB. The estimations had a large between-subject variability with a length of the interquartile ranges (IQRs) between 3.25 dB (pair 57) and 19.0 dB (pair 23) with a median value of 8 dB. These results were 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 and are being 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 sessions 1 and 2 the same six 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

![FIG. 4. (Color online) Discrimination thresholds for the instrument-in-noise tests using (A) non-reverberant sounds and (B) reverberant sounds. The thresholds (red triangles in (A) and green squares in (B) are used as a measure of similarity between the sounds, and were assessed taking the median across participants. The piano pairs are shown along the abscissa and were ordered from most to least similar (higher to lower SNR thresholds). The error bars represent interquartile ranges (IQRs).)](image-url)
thresholds of session 1. From the 210 obtained thresholds, 31 estimations were excluded.

1. Exclusion criteria

Thirty-one staircases were excluded from the data analysis after the data collection. Three staircases were incomplete with less than 12 reversals. Another three staircases were removed because the participants reached a pre-established maximum SNR of 50 dB (minimum noise level). In such a case the participants were not able at all to discriminate the test sounds. 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 controlled 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 Fig. 5 where one of those staircases is shown. The staircase has a convergence point (last four reversals) that differs from the threshold estimation by 3.5 dB.

2. Thresholds using a constant stimulus procedure

As part of our hypotheses the discrimination task at high SNRs should have been easy with scores of nearly 100%. This was not the case for pair 47 (and pair 74), where two staircases obtained from the first five participants had to be excluded according to the criteria described above. For this reason, the remaining participants were asked to conduct 16 3-AFC trials at a fixed SNR of 20 dB, obtaining scores of 81.25%, 50.0, 81.25%, 50.0%, and 68.75%. These scores were converted into SNR thresholds (70.7%-points), assuming a performance slope of 4.1%/dB,8 obtaining thresholds of 17.5, 25.0, 17.5, 25.0, and 20.5 dB, respectively. These scores the results of each triad, two points were attributed to the pair indicated as most similar, no points to the least similar pair, and one point to the remaining pair. Since each sound pair was tested 5 times by 20 participants, the maximum possible score $S_{\text{max}}$ was 200. The assessed scores ranged between $S_{ij} = 33$ (pair 24) and 190 (pair 23) as shown in Table II.

The similarity scores $S_{ij}$ were then converted into a measure of dissimilarity by using

$$D_{ij} = \sqrt{1 - S_{ij}/S_{\text{max}}}.$$

with $S_{ij}$ being each element of the similarity matrix, $S_{\text{max}} = 200$, and $D_{ij}$ are the elements of the new dissimilarity matrix. The dissimilarity matrix was then used as input to the classical MDS algorithm (Everitt, 2005) available in the MATLAB Statistics Toolbox. The MDS algorithm returns a Cartesian space with $q$ dimensions containing the $n = 7$ test stimuli, where $q < n$. As criterion to test the adequacy of the $q$-dimensional representation we assessed the residual sum of squares between the dissimilarity scores $D_{ij}$ and the Euclidean distances $d_{ij}$ obtained from the fitted space. This calculation leads to a number that is referred to as stress $S_r$ (Kruskal, 1964b) and is given by

$$S_r = 100 \sqrt{\frac{\sum_{i<j} (D_{ij} - d_{ij})^2}{\sum_{i<j} D_{ij}^2}}.$$
where the lower the $S$-value, the better is the goodness-of-fit of the $q$-dimensional space. The lowest possible stress is $S = 0\%$, which indicates that $d_{ij}$ and $D_{ij}$ have a perfect monotone relationship.

The resulting space had $q = 4$ dimensions with a goodness-of-fit $S = 3.1\%$ (good to excellent) with cumulative stresses of 21.9% (poor) for the first two dimensions and 7.5% (fair to good) for the first three dimensions. The first two space dimensions ($S = 21.9\%$) are depicted in Fig. 6(A), where the location of each piano is shown along with bubbles indicating the variability in responses across participants. Although this reduced representation provides a poor fit ($S > 20\%$), 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 distances $d_{ij}$ between pianos are shown in the lower left triangle of Table II and they are indicated as filled square markers in Fig. 7(A). The distances $d_{ij}$ ranged between 0.14 (pair 47) and 0.91 (pair 24) with an IQR between $d_{ij,25} = 0.63$ and $d_{ij,75} = 0.83$.

The results shown in Fig. 6(A) suggest that the non-reverberant sounds (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 seemed to have an intermediate similarity with all these groups, in the four-dimensional space its distances increased systematically. The distances for all the other pianos did not differ considerably with respect to the ones in the two-dimensional representation.

### 2. Between-subject variability

An estimate of the variability in responses across participants is given by the diameter of the bubbles in Fig. 6(A), with a smaller diameter denoting more consistency in the participants’ judgment for the corresponding piano.

To assess this variability, five new MDS spaces were generated from the experimental data of participants S01–S04, S05–S08, S09–S12, S13–S16, and S17–S20. This resulted in five new coordinates for each of the seven pianos (P01–P07). For each of the seven pianos, the distance between the five new coordinates and the global space was calculated, storing the difference between the minimum and maximum distances. This difference was eventually used as diameter of the bubbles in Fig. 6(A).

The diameters ranged between 0.06 (piano P4) and 0.29 (piano P5) with a median of 0.15. The interpretation of this is that piano P4 was judged more consistently across participants and, correspondingly, piano P5 was judged more differently with a higher between-subject variability. The obtained five four-dimensional spaces were used to assess the minimum and maximum distances between piano pairs, which are shown as error bars in Fig. 7(A). Those deviations ranged from 0.05 (pair 57) to 0.33 (pair 16) with a median length of 0.17.

### C. Dataset 2: Instrument-in-noise test

The discrimination thresholds obtained for reverberant piano sounds are shown in Fig. 4(B). The pooled thresholds (green square markers) were assessed by taking the median of

![FIG. 6. (Color online) Perceptual MDS spaces for (A) non-reverberant, and (B) reverberant piano sounds. Only the first two dimensions are shown. Although the goodness-of-fit of these reduced representations is poor (dataset 1: $S = 21.9\%$; dataset 2: $S = 29.2\%$) the overall distribution of the pianos did not change in the corresponding four-dimensional spaces. The bubbles give an indication of the participant’s variability. Refer to the text for further details.

![FIG. 7. (Color online) Euclidean distances $d_{ij}$ taken from the corresponding four-dimensional MDS space for (A) non-reverberant and (B) reverberant piano sounds. The $d_{ij}$ values are also shown in the lower left triangle of Table II (across columns P1–P7 for dataset 1 and P1–P7, for dataset 2). The piano pairs are sorted in the same way as in Fig. 4. The error bars indicate the minimum and maximum distances between piano pairs across five MDS spaces obtained by grouping the experimental data in smaller subsets (see the text for further details).]
all threshold estimations per piano pair. The thresholds ranged between 24.25 dB (pair 47) and −4.0 dB (pair 24), covering a SNR-range of 28.25 dB. The estimations had a large between-subject variability with IQRs between 5.0 dB (pair 24) and 16.6 dB (pair 46) with a median value of 11.0 dB. These results were based on 189 staircase threshold estimations. From the 210 obtained thresholds, 21 estimations were excluded.

1. Exclusion criteria

Twenty-one staircases were excluded from the data analysis after data collection. Seven staircases were removed because the participants reached a pre-established maximum SNR of 50 dB. The remaining 14 thresholds were removed after a check of consistency of the staircases. For this the standard deviation of the reversals was assessed. Threshold estimations where the deviation of the reversals was larger than 4 dB were removed. This criterion is less strict than the criterion used for non-reverberant sounds, which was based on a deviation of 3 dB. If this criterion would have been adopted, 24 other staircases should have been excluded (total of 45 exclusions, 21% of the data). We decided to set the criterion to 4 dB to preserve more data points.  

D. Dataset 2: Triadic comparisons

1. Construction of the similarity matrix

The results of all participants (S01–S20) were used to construct the similarity matrix $S_{ij}$ shown in the upper triangle across the columns labeled as $P1_r$–$P7_r$ in Table II. The assessed scores ranged between $S_{ij} = 28$ (pair 24) and 183 (pair 47).

2. MDS

The triadic comparison data were processed by first converting the similarity scores $S_{ij}$ into counts of dissimilarity $D_{ij}$ [Eq. (1)]. The dissimilarity matrix was then used as input to the non-metric MDS algorithm available in the MATLAB Statistics toolbox. An a priori number of $q = 4$ dimensions (as obtained from dataset 1) was used to obtain the perceptual space. The resulting space had a stress $S_r$ of 6.9% (close to good) with cumulative stresses of 29.2% and 12.7% for the first two and three dimensions, respectively. The distances $d_{ij}$ of the fitted space are shown in the lower left triangle of Table II across the columns labeled as $P1_r$–$P7_r$. The first two space-dimensions are shown in Fig. 6(B). The results in this reduced representation suggest that the reverberant sounds can be classified in five distinct groups: pianos $P2_r$ + $P3_r$, $P4_r$ + $P7_r$, $P1_r$, $P5_r$, and $P6_r$. We labeled piano $P6_r$ as having intermediate similarity with $P4_r$ and $P7_r$ despite their overlapped position in the two-dimensional space. This is due to the relative change of location of $P6_r$ when adding dimension three of the space (not shown here).

3. Between-subject variability

The same processing scheme applied to the non-reverberant data to assess between-subject variability was used in the analysis of the reverberant piano data (see Sec. IV B 2), resulting in the bubbles depicted in Fig. 6(B). The diameter of the bubbles ranged between 0.06 (piano $P3_r$) and 0.22 (piano $P5_r$) with a median of 0.14. This means that piano $P3_r$ was more consistently judged across participants while piano $P5_r$ was scored with more variability. The obtained spaces were also used to assess the minimum and maximum distances between piano pairs, which are shown as error bars in Fig. 7(B). Those deviations ranged between 0.03 (pair 26) and 0.30 (pair 16) with a median length of 0.18.

V. DISCUSSION

A high perceptual similarity is equivalent to a high SNR threshold and a short Euclidean distance $d_{ij}$. If the results of both methods are consistent, the SNR thresholds of Fig. 4, which are sorted in decreasing order, should correspond to monotonically increasing distances $d_{ij}$, and a perfect consistency between methods would be given by a correlation value of −1. Both Pearson and Spearman (rank-order) correlations are reported next. Our focus is, however, on rank-order correlations given that it was a priori unclear whether the two test measures should be linearly related, considering that one measure represents dimensionless distances ($d_{ij}$), and the other measure is expressed in a logarithmic scale (SNR).

A. Dataset 1: Comparison between methods

For non-reverberant piano sounds the results of both experimental methods, instrument-in-noise test and triadic comparisons, were significantly correlated with a moderate to high (Pearson) correlation $r_{p}(17) = −0.47$, $p = 0.04$, and a high rank-order (Spearman) correlation $r_{s}(19) = −0.65$, $p = 0.001$. The scatter plot and linear regression analysis of these data are shown in Figs. 8(A) and 8(C), where an overall inverse relationship between variables is observed. Further inspection of the data [Figs. 4(A) and 7(A)] revealed that the two most similar pairs are the same in both methods (pairs 23 and 47). Furthermore, the methods coincide in the judgment of three of the six most different pairs (SNR25–75 > 0.5 dB, distances $d_{ij} > 0.8$): 26, 27, 37. Piano $P5$ had an intermediate similarity with the other piano sounds with distances $d_{ij}$ between 0.56 (pair 56) and 0.78 (pair 15), this means that 5 (out of 6) distances were within the IQR of the distance data ($d_{(0.25-0.75)} = 0.63–0.83$). This is also supported by the results of the instrument-in-noise test, where five (out of six) thresholds were within the IQR (SNR25–75 = 0.2 – 7.7 dB). For two pairs (16 and 56), the methods provided very different similarity measures. In both cases, the pairs were judged as being more similar in the triadic comparisons.

B. Dataset 2: Comparison between methods

For reverberant piano sounds the results of both methods, instrument-in-noise and triadic comparisons, were significantly correlated with a moderate to high (Pearson) correlation $r_{p}(18) = −0.50$, $p = 0.03$, and a high rank-order correlation $r_{s}(19) = −0.65$, $p = 0.001$. The scatter plot and linear regression analysis of these data are shown in Figs. 8(B) and 8(D), where an overall inverse relationship between variables is
observed. Further inspection of the data [Figs. 4(B) and 7(B)] revealed that the methods share two of the three most similar pairs (pairs 47 and 36). Furthermore, the methods coincide in the judgment of three of the five most different pairs (SNR thresholds \(< 1.9 \text{ dB}, d_{25} > 0.80\): 24, 26, 27. There were, however, some pairs for which the methods provided different similarity measures. If the IQRs of the results are used to delimit three similarity regions (high, \(d_{25} < 0.63, \text{thres}_{75} > 6.6 \text{ dB};\) medium, \(d_{25} > 0.63, \text{thres}_{75} < 6.6 \text{ dB}\), then five piano pairs were judged differently:

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

The larger number of differing judgments may be due to the apparent increase in difficulty of the tasks with respect to the use of dataset 1. Evidence for this is: (1) the lower stresses of the MDS space with respect to the similarity matrix \((S_{\text{rev}} = 6.9\% \text{ in contrast to } S_{\text{non-rev}} = 3.1\%\) and the poorer cumulated stresses for the first two and three dimensions \((S_{\text{rev}} = 29.2 \text{ and } 12.7\% \text{ compared with } S_{\text{non-rev}} = 21.9 \text{ and } 7.5\%, \text{ respectively})\), and (2) the larger number of excluded staircases if the same criterion as for dataset 1 would have been adopted. Despite the apparent increase in the difficulty of the tasks, similar correlations between methods were found using both datasets \([r_s(19) = -0.65, p = 0.001]\).

**C. Comparison among sounds without and with reverberation**

To give some idea of the effect that reverberation has on the similarity results in the triadic comparison and instrument-in-noise methods, rank-order correlations between the non-reverberant and reverberant Euclidean distances \(d_{ij}\) (Table II), as well as with the corresponding thresholds (Fig. 4), were assessed. The distances \(d_{ij}\) obtained from dataset 1 and dataset 2 were significantly correlated with \(r_s(19) = 0.81, p < 0.001\). Similarly, the SNR thresholds exhibited a significant, although slightly lower correlation with \(r_s(19) = 0.70, p < 0.001\). For the latter case, the sounds did not only differ in terms of reverberation but they also differ in the slightly different spectral weighting of the ICRA noises (versions A and B). It might be that the spectral tilt, which was briefly discussed in Sec. II, was responsible for the lower rank-order correlation between thresholds. Based on our rationale of accurately following the spectral and temporal properties of the test (piano) sounds, the ICRA algorithm version B should be preferred in future tests rather than version A.

**VI. CONCLUSION**

In this paper we have presented a method to assess the perceptual similarity of two sets of piano sounds using an instrument-in-noise test. The noise used in the test follows the spectro-temporal properties of the sound pairs being tested.
A. Similarity among 19th-century Viennese pianos

As a study case, a comparison among recordings of one note (C♯₅) played on seven 19th-century Viennese pianos was conducted. The sounds were compared in two conditions: (1) “as recorded” in a condition assumed to be nearly anechoic (non-reverberant sounds, dataset 1), and (2) by adding the effect of reverberation of a room that had an EDTₘₐₐₜ of 3 s (reverberant sounds, dataset 2). The instrument-in-noise test was compared to the method of triadic comparisons. The results of both methods, collected from two different groups of 20 participants, were significantly correlated with rank-order correlations of −0.65 for both datasets (Fig. 8). These correlation values denote a high inverse correlation between SNR thresholds and Euclidean distances, meaning that (on average) higher thresholds correspond to short Euclidean distances. The instrument-in-noise method is therefore correlated with subjective similarity judgments and seems to be a promising method to quantify perceptual differences between sounds.

B. Modifying the perceptual similarity

It was pointed out that the instrument-in-noise method is rather time consuming, requiring about 3 h to evaluate the same set of sounds that could be evaluated in less than 30 min using triadic comparisons. However, one of the primary advantages that we see in the suggested method is that it allows to measure similarity in conditions (different SNRs) where the physical properties of the test sounds are affected. The use of different noise levels represents, therefore, a quantifiable way to manipulate the similarity between 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) that leads to purely psychological (similarity) judgments, where the physical sound properties are kept constant throughout the evaluation. With this argument, the instrument-in-noise method can give an indication not only of which sounds are closer or farther apart from one another (psychological approach), but can also provide evidence about their acoustic properties at noise levels below and at thresholds (physical approach).

C. Extending the use of the instrument-in-noise test

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) was adapted to provide a suitable solution for instruments sounds. The 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 considered. For our piano sounds, these aspects were: to have test stimuli with the same pitch, similar durations, a piano onset occurring at a “synchronized” time stamp, and to balance for any cue that is not desired to be judged (e.g., loudness and pitch). Some 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 musical instruments, noises have to be generated again to match the spectro-temporal properties of the new target sounds. It should be noted that with the rationale of accurately following the properties of the target sounds, the ICRA noise algorithm version B should be used in future evaluations.

ACKNOWLEDGMENTS

This research work was funded by the European Commission (EC) within the ITN Marie Skłodowska-Curie Action project BATWOMAN under the Seventh Framework Programme (EC Grant No. 605867) and was further supported by the Lise-Meitner Fellowship M1653-N30 of the Austrian Science Fund (FWF) attributed to A.C.

1In this study we use the definition of triadic comparisons as presented by Levelt et al. (1966). An alternative method, which has also been named “triadic comparisons,” is given by Wickelmaier and Ellermeier (2007). Although the latter method is also based on comparisons of sound triads, it fundamentally differs from the method described by Levelt et al. in that rather than quantifying perceptual similarity, its purpose is to determine a qualitative feature structure for the stimulus set under study.

2By averaging two waveforms, the variance of the resulting paired noise is actually decreased by 3 dB.

3Despite the fact that loudness is an intrinsic property linked to the evolution of the piano construction, the timbre of the resulting sounds is also affected. Since the timbre characteristics of the piano sounds were not manipulated, there are (partly) discrimination cues left after the loudness balancing process.

4The RIR used in the auralization corresponds to an existing measurement of the former church Aula Carolina (Aachen, Germany), which has a ground area of 570 m² and a high ceiling. The selected RIR was measured at a distance of 4 m and azimuth of 90° with respect to the sound source, and it was retrieved from the AIR database (Jeub et al., 2009).

5With the exception of one participant, aged 66 yr, all participants were between 19 and 26 years of age at the time of testing. Their hearing thresholds were not measured, but we assumed a normal hearing condition. The participant aged 66 years, however, may have had some hearing loss, but because all his staircases met at least one of the exclusion criteria, his data were not further used.

6For instance, if pair 57 (i.e., P5 being the target sound and P7 being the reference sound) was attributed to one participant, then pair 75 (P7 being the target sound and P5 being the reference sound) was attributed to the other participant.

7During the experimental pilots no evident effect of piano order was found. For this reason we treated pairs AB and BA as being equivalent comparisons. When dividing the obtained discrimination thresholds [Fig. 4(A)] into AB and BA pairs the 21 pair differences had a signed average value of −0.9 dB and an unsigned value of 4 dB.

8The performance rate of 4.1% was obtained for one participant (S06), who tested pair 47 using both the adaptive and constant stimulus procedures. The participant had an estimated threshold of 23.5 dB (assumed score of 70.7%), and the score obtained at 20 dB SNR was 56.25%, 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 spite of the lack of experimental evidence for this assumption, simulated thresholds obtained by feeding the same piano sounds into a computational model (Osses and Kuhlrausch, 2018) 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).

9This method to derive the variability in the experimental responses is similar to the solution provided by the individual differences scaling algorithm (INDSCAL; Carroll and Chang, 1970). A substantial difference is that within INDSCAL a new space is obtained using each participant’s data with the assumption that each of individual space is a weighted version of a common (global) space. We did not adopt this approach because the obtained global space violated the condition of monotonicity between \( D_{ij} \) and \( d_{ij} \).
Although not shown here, the overall thresholds did not change significantly by adopting either criterion (3 or 4 dB). The thresholds $\text{thres}_{5-dBcrit}$ and $\text{thres}_{4-dBcrit}$ (as finally used for reverberant piano sounds) had a correlation $r(19) = 0.93, p < 0.001.$


