Computational Models of Similarity for Drum Samples

Elias Pampalk, Perfecto Herrera, and Masataka Goto

***Abstract*—In this paper, we optimize and evaluate computational models of similarity for sounds from the same instrument class. We investigate four instrument classes: bass drums, snare drums, high-pitched toms, and low-pitched toms. We evaluate two similarity models: one is defined in the ISO/IEC MPEG-7 standard, and the other is based on auditory images. For the second model, we study the impact of various parameters. We use data from listening tests, and instrument class labels to evaluate the models. Our results show that the model based on auditory images yields a very high average correlation with human similarity ratings and clearly outperforms the MPEG-7 recommendation. The average correlations range from 0.89–0.96 depending on the instrument class. Furthermore, our results indicate that instrument class data can be used as alternative to data from listening tests to evaluate sound similarity models.**

***Index Terms*—Content-based similarity, drum sounds, percussive music instruments.**

I. Introduction

DRUM samples are used to create drum loops, which play an important role in several styles of music. Large commercial drum sample libraries can contain thousands of samples. Computational models of similarity enable new user interfaces which allow music producers to find and retrieve samples more easily. For example, such an interface could automatically organize drum samples hierarchically and visualize them on a two-dimensional map such that similar samples are located close to each other [1]. An alternative application is to retrieve a set of similar samples given a query sample.

In contrast to most of the previous work, we are not focusing on classifying the samples according to instruments or comparing different instrument classes. Instead, we want to compute the similarities of samples from the same instrument class. We investigate four instrument classes: bass drums, snare drums, high-pitched toms, and low-pitched toms. We limit our scope to dry drum sounds.[1] Two samples from the same class differ, for example, with respect to how hard the instrument was hit, the diameter of the drum, the manufacturer, or the recording environment.

We evaluate two distinct similarity models. The first model is closely related to work on timbre spaces and is part of the MPEG-7 standard.[2] The second model is based on aligned auditory images. To evaluate the models, we use data from two sources. First, we use human similarity ratings gathered in listening tests. For each of the four instrument classes, we conducted a listening test where subjects were asked to rate the similarity of pairs of sounds. Second, we use the instrument class labels of a collection with 311 drum samples. Our results show that the model based on auditory images fits the data with a significantly lower error than the MPEG-7 model. Furthermore, our results show that instrument class data can be used (at least in early development stages) to replace data gathered in listening tests.

The remainder of this paper is structured as follows. In Section II, we review related work. In Section III, we describe the computational models of similarity. In Section IV, we describe the data we use for the evaluations and analyze it. In Section V, we present the evaluation of the models. Finally, in Section VI, we summarize our findings and point out directions for future work.

II. Related Work

There is a vast amount of work on the similarity of sounds from music instruments. In this section, we briefly review work on timbre spaces, auditory images, instrument classification, content description and MPEG-7, and finally, the larger context and applications.

A. Timbre Spaces

Timbre is the quality that distinguishes sounds with the same pitch, loudness, and duration.[3] The perception of timbre is multidimensional, subjective, and context dependent. Work on timbre spaces assumes that there is a low-dimensional space (often three-dimensional) in which similarities can be explained. Each of these dimensions is associated with measurable physical properties of the audio signals which can be extracted automatically.

Manuscript received December 15, 2006; revised September 16, 2007. This work was supported in part by the SIMAC project and the CrestMuse project (CREST, JST). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gaël Richard.

E. Pampalk is with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568, Japan. Part of this work was done while E. Pampalk was with the Austrian Research Institute for Artificial Intelligence (OFAI). (e-mail: elias.pampalk@aist.go.jp).

P. Herrera is with the Music Technology Group, University Pompeu Fabra, 08003 Barcelona, Spain (e-mail: perfecto.herrera@iua.upf.es).

M. Goto is with the Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568, Japan (e-mail: m.goto@aist.go.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.910783

[1]Dry sounds are sounds to which no additional effects such as reverberation have been added.

[2]ISO/IEC 15938 Information technology—Multimedia content description interface—Part 4: Audio (2002)

[3]American Standards Association (1960). *American Standard Acoustical Terminology*. New York. Definition 12.9, Timbre, p. 45.

The dimensions are found and linked to physical properties by analyzing human listeners' similarity ratings. These ratings are gathered in listening tests where subjects are asked to rate the pairwise timbre similarity of sounds. One approach to analyze these pairwise similarities is to visualize the space they define using multidimensional scaling (MDS).[4] The first study using MDS to analyze the dimensions of timbre was published by Grey [3]. Grey developed a model with three dimensions, where the most prominent one is related to the distribution of the spectral energy. This research direction was continued in a number of studies (see, e.g., [4]–[7]).

B. Auditory Images

Auditory images have the dimensions: time, frequency, and loudness. For the frequency and loudness dimensions, a nonlinear perceptual scale is used. The similarity of sounds is computed by comparing their auditory images. Work related to auditory images includes [8]–[15]. The approach we use is mostly based on [1] and [8].

Closely related to work on auditory images is work on measuring the perceptual audio quality (e.g., [16]–[18]), where the objective is to measure very small differences between two sound excerpts (for example, to evaluate lossy encoders).

In general, there is a vast amount of related work regarding the perception of loudness. For an engineer's perspective, see, e.g., [19]. For an approach which models the auditory system down to the level of the mechanical response of the hair cells and the electrical response of auditory nerve fibers, see, e.g., [20].

C. Instrument Classification

The goal of instrument classification is to identify the instrument class of a given sound. The sounds can either be isolated (without any background noise) or extracted from a piece of music (and thus include other instruments and voices in the background). When a nearest neighbor classifier is used to classify the sounds, the relationship to similarity models is most obvious. However, even if more complex classifiers are used, there is a lot of overlap with our work. First, features which help classify instruments, such as the spectral centroid or the attack time, are very likely to also be useful when computing similarities. Second, as we show in Subsection V-B instrument, class data can be used to evaluate similarity models.

An extensive overview of related work on automatic classification of pitched musical instrument sounds can be found in [21] and, for unpitched sounds, in [22]. In this subsection, we only mention some examples with a particular focus on the features used.

One of the first publications on audio classification was [23], where class models of sounds (mostly "sound effects") were built using loudness, pitch, and spectral descriptors such as the centroid or the bandwidth, in addition to their variation along time. The list of features was extended in [24] to deal with the classification of orchestral sounds. Work on distinguishing between oboe and sax sounds using only cepstral coefficients (without temporal information) was presented in [25]. The results show that the proposed system performs about as well as trained human listeners. The results published in [26] show that there are several sets of features which can be used to achieve performances similar to those of humans.

One of the first approaches focusing on percussion sounds is [27]. The results showed that very simple features (in particular the zero-crossing rate) are very useful to discriminate instruments. Further work on drum sounds includes [28], where classification accuracies are reported of around 90% for nine different instrument categories: bass drum, snare drum, tom (low, medium, high), hihat (open, closed), and cymbal (ride, clash). Furthermore, the results showed that it is possible to reduce the original feature set of size 50 to about 20 without performance loss. The most powerful features found were: zero crossing rate, and spectral shape descriptors (kurtosis, skewness, centroid, low-order Mel frequency cepstral coefficients), and relative energy in low-frequency bands. A larger study was conducted in [29] with 208 different features, 30 different unpitched percussive sound categories, and different classifiers. In addition, MDS was used to visualize the feature space and understand errors of the classification system.

Recently, research has focused on describing or classifying sounds produced by a single instrument. Work focusing on the clarinet and studying temporal variations using self-organizing maps (SOMS) was presented in [30]. Work focusing on the shakuhachi was presented in [31], oboe sounds in [32], and drum sounds and different playing modes in [33]. Overall, Mel frequency cepstral coefficients (MFCCs) have been the most prevalent and effective features for addressing problems of musical sound classification.

D. Content Description and MPEG-7

In the context of the MPEG-7 standard there have been a number of publications on automatically extracting semantic descriptors from audio signals. Of particular interest for this work is the work presented in [34]. We have implemented this MPEG-7 model for the distance of percussive sounds and compared it to a model based on auditory images.

E. Larger Context and Applications

On a larger scale, related work includes, for example, identifying instruments within polyphonic audio (e.g., [35] and [36]). Another example is the work presented in [37], where drum sounds are identified within a piece of music using templates similar to auditory images. Once the sounds are identified, rhythm patterns can be extracted, or various effects can be applied (e.g., [38]). Other related topics include, for example, mapping semantic meanings to features extracted from the audio signal (e.g., [39]). More direct applications include, for example, the hierarchical organization of drum sample libraries [1]. Such organizations can help producers creating and tweaking drum loops find the samples they are searching for more easily.

III. Computational Models

We compare two distinct models. The first model is part of the MPEG-7 standard which is closely related to work on timbre

[4]MDS is a technique to project a high-dimensional space to a lower dimensional space while preserving (as far as possible) the distances between the data [2].

spaces as discussed in Subsection II-A. The second model is based on auditory images. For both models we only used the left channel (mono) of each sample (sampled at 44.1 kHz) as input.

A. Timbre Space (MPEG-7 Standard)

The MPEG-7 standard[2] defines a method to compute the similarity of percussive sounds (including drum sounds) which is based on work presented in [34]. The three dimensions that are measured for each sound are log-attack time (LAT), temporal centroid (TC), and spectral centroid (SC). Generally, a drum hit very hard has a shorter attack time. The temporal centroid depends on factors such as the resonance (or dampening) and reverberation. The spectral centroid is related to the brightness. The following paragraphs describe how they are defined according to the MPEG-7 standard.

Let $S(n)$ be an approximation of the signal's power function over time, where n is the index of the time frame, N is the total number of frames, and sr is the sampling rate (in Hertz). $S(n)$ is computed as the local mean square value of the signal amplitude within a 25-ms running window.

The log-attack time is defined as

$$\mathrm{LAT} = \log 10(t_1 - t_0) \tag{1}$$

where t_0 is the point in time at which the signal power exceeds 2% of the maximum value, and t_1 is the point at which the signal power level reaches peak level (both measured in seconds).

The temporal centroid is defined as

$$\mathrm{TC} = \frac{\sum_{n=1}^{N} \frac{n}{sr} S(n)}{\sum_{n=1}^{N} S(n)}. \tag{2}$$

The spectral centroid is calculated from the power spectrum of the whole sample and is defined as

$$\mathrm{SC} = \frac{\sum_{k=1}^{K} f(k) C(k)}{\sum_{k=1}^{K} C(k)} \tag{3}$$

where $C(k)$ is the kth power spectrum coefficient, $f(k)$ is the frequency of $C(k)$, and K is the total number of coefficients.

Given two samples A and B, their distance is computed as

$$d_{AB} = \sqrt{\left(d_{\mathrm{LAT}} \frac{w_1}{10} + d_{\mathrm{TC}} \frac{w_2}{10}\right)^2 + \left(d_{\mathrm{SC}} \frac{w_3}{10^5}\right)^2} \tag{4}$$

where $d_{\mathrm{LAT}} = \mathrm{LAT}_A - \mathrm{LAT}_B$ and so forth. In the MPEG-7 standard the following examples are given for the coefficients: $w_1 = 3$, $w_2 = 6$, $w_3 = 10$. However, these specific values for the coefficients are provided for informative purposes only. The standard explicitly states that these coefficients may not be appropriate for arbitrary data sets. In the experiments described later we use a grid search to optimize these coefficients.

B. Auditory Images

Auditory images have the dimensions time (columns), frequency (rows), and loudness (values). Each image can be interpreted as a very high-dimensional vector (e.g., by concatenating the rows). These vectors can directly be compared to each other by computing the Euclidean distance (or the Minkowski distance of any order). However, this approach is suboptimal if the sounds are not temporally aligned (as seen in Fig. 1). In our work, we use a simple brute force approach where we try all possible alignments within certain limits and select the one which fits best [1]. An alternative approach, for example, is to use an approach based on matching trajectories computed using a self-organizing map [10].

The following subsections describe how we compute the auditory images. These computations are influenced by nine parameters which we optimize in the next section.

1) Sample Length: In a first step, we consider using only a certain amount of the beginning of the signal. In particular, we consider using only the first 250 ms, 1000 ms, or the whole signal. Using only 250 ms puts an emphasis on the onset. While the long tail might be hardly audible, it can still have a significant impact on the distance computation. As we discuss later the improvements using only the first 250 ms for the bass drum (BD) sounds indicate that there is a potential for techniques which put more weight on the perceptually relevant parts of the sample. This confirms the findings presented in [13], where the correlation of the subjects' ratings and the auditory image based model increased from about 0.68 to 0.88, mainly due to an increased focus on the onset.

2) Power Spectrum: Given the audio signal, the power spectrum is computed with a STFT using a Hann window function. The window size and the overlap between windows are two critical parameters. In our experiments, we evaluate the window size parameter for values in the range of 256 samples (6 ms) to 16 384 samples (372 ms). The overlap is defined in terms of the relative hop size of the moving window. We evaluate the range from 1 (no overlap) to 1/8 (87.5% overlap).

3) Outer and Middle Ear Response: A model for the outer and middle ear frequency response can be applied optionally. In particular, we evaluate if application of the filter suggested by Terhardt [40] improves the similarity model. The response of the filter is defined as,

$$A_{\mathrm{dB}}(f_{\mathrm{kHz}}) = -3.64\, f^{-0.8} + 6.5 \exp(-0.6(f-3.3)^2) - 10^{-3} f^4. \tag{5}$$

The main characteristics of this filter is that it reduces the impact of very high and very low frequencies. On the other hand, frequencies around 3–4 kHz are emphasized.

4) Frequency Scale: The human perception of frequency is not linear. Among the various models that have been proposed (see, e.g., [41]), we focus only on three. The simplest model we use is logarithmic scaling (with dual basis). Second, we use the Mel scale which is also used to compute MFCCs. Third, we use the Bark scale, which despite having a very different background has very similar characteristics to the Mel scale.

The Mel scale [42] is defined as

$$Z_{\mathrm{mel}}(f_{\mathrm{kHz}}) = \ln(1 + f/0.7) * 1127.01048. \tag{6}$$

The Bark scale (see, e.g., [43]) is defined as

$$Z_{\mathrm{bark}}(f_{\mathrm{kHz}}) = 13 \arctan(0.76 f) + 3.5 \arctan(f/7.5)^2. \tag{7}$$

The main characteristic of the Bark scale is that the width of the critical-bands is about 100 Hz up to 500 Hz, and beyond 500 Hz the width increases nearly exponentially. The Mel scale has very similar characteristics. The main difference between the two scales is that the Mel scale was developed by measuring when a tone is perceived to have a pitch twice as high (or low) as a reference tone. The Bark scale was developed by measuring the differences in frequencies when spectral masking occurs (see the next subsection).

The frequency scale defines the center frequencies and widths of the frequency bands we use. For all three scales, we use triangular filters to compute the energy of each frequency band. The triangles overlap such that the upper frequency of one triangle is the peak frequency of the next triangle and the lower frequency of the one after the next. Such triangular filters are generally also used to compute MFCCs. The implementation we used for the Mel scale is available online [44].

The frequency for the triangle with the lowest center frequency is fixed at 10 Hz, the frequency for the triangle with the highest center frequency is fixed at 13 500 Hz. (The center frequencies of all other bands vary depending on the number of frequency bands used.) The two parameters we evaluate are the frequency scale and the number of frequency bands (in the range of 6 to 72).

5) Spectral Masking: Spectral masking is the occlusion of a sound by a louder and simultaneous sound with a similar frequency. Applying the spectral masking model is optional. We compute the masking effects according to Schroeder *et al.* [45] who suggest a spreading function optimized for intermediate speech levels. Alternatively, more complex models could be used [46].

The spreading function suggested by Schroeder *et al.* has lower and upper skirts with slopes of $+25$ dB and -10 dB per critical-band (according to the Bark scale). For the Mel and logarithmic frequency scales, we compute the critical-band equivalent using the center frequencies of the respective frequency bands.

The main characteristic of the masking function is that lower frequencies have a stronger masking influence on higher frequencies than vice versa. The contribution B_{dB} of critical-band z_i to z_j is computed by

$$\begin{aligned} &B_{\mathrm{dB}}(\Delta z_{\mathrm{bark}}) \\ &\quad = +15.81 + 7.5(\Delta z + 0.474) + \\ &\quad - 17.5(1 + (\Delta z + 0.474)^2)^{1/2} \end{aligned} \tag{8}$$

where $\Delta z = z_j - z_i$. In general, applying the spreading function can be interpreted as asymmetric smoothing of the frequency resolution. The effect of this smoothing depends on the number of frequency bands used. In particular, applying the spectral masking model will show hardly any effects if only few bands are used (because Δz increases). Furthermore, not only the number of frequency bands but also the number of discrete cosine transform (DCT) coefficients used (described later on) are closely related to spectral masking.

6) Loudness: Loudness is the third dimension of the auditory images. The perception is not linear with respect to the energy in the audio signals. A popular approximation is to apply a decibel (dB) scaling to the energy. In addition, we evaluate a loudness model suggested in [47]. Given the sound pressure level in dB, the loudness in sone is computed as

$$S_{\mathrm{sone}}(l_{\mathrm{dB}}) = \begin{cases} 2^{(l-40)/10}, & \text{if } l \geq 40 \text{ dB} \\ (l/40)^{2.642}, & \text{otherwise.} \end{cases} \tag{9}$$

7) DCT Compression: A DCT can be applied to obtain a compressed representation of each time frame of the auditory image (a vector of loudness values with as many dimensions as frequency bands). Computing the DCT is a standard step for MFCCs.

The DCT transformation describes the original frame as a linear combination of orthogonal sinusoids. High-frequency sinusoids are omitted to obtain a compressed representation. In addition to computational advantages, this compression can also be interpreted as a form of spectral masking. In particular, omitting high-frequency components leads to stronger smoothing along the frequency scale.

8) Interpretation as Images: Up to this point, we have obtained a representation for each sound in the dimensions: frequency, time, and loudness. The scale and resolution of each dimension depends on parameters. For example, Mel scaled frequency and loudness in dB are the standard approach when computing MFCCs. The resolution on the frequency scale depends on the number of frequency bands, the resolution on the (linear) time scale depends on the STFT hop size. We refer to this representation as an auditory image (see Fig. 1 for examples). Time is mapped to the horizontal dimension, and frequency to the vertical dimension. The loudness dimension is mapped to the "color." Given a specific set of parameters, the images always have the same height. However, their width depends on the length of the audio signal.

9) Alignment and Distance Computation: A simple approach to compare two images is the following. If they are not the same length, then the shorter one can be appended with zero columns at the end (which correspond to silence). Each image can be interpreted as vector by concatenating the rows. These vectors can be compared using any distance metric applicable to vector spaces such as the Euclidean distance or the Minkowski distance of any order.

As can be seen in Fig. 1, the samples are not always temporally aligned. However, the alignment has a large impact on the computed distance using the simple approach described above. Thus, it is necessary to align the images before computing the distance.

We apply the following approach to align two images A and B. We shift A (on the time axis) against B (within reasonable limits). This is done by inserting or appending silence. For every possible shift we compute the distance between the shifted version of A and B. We define d_{AB} to be the minimum distance from all temporal alignments.

The step size of the temporal shifts is defined by the temporal resolution. We shift the images in the range from 0 to 100 ms. Simple heuristics can be applied to minimize the number of necessary computations such as aligning the images using only the sum over all frequency bands, or aligning all samples with a prototypical sample.

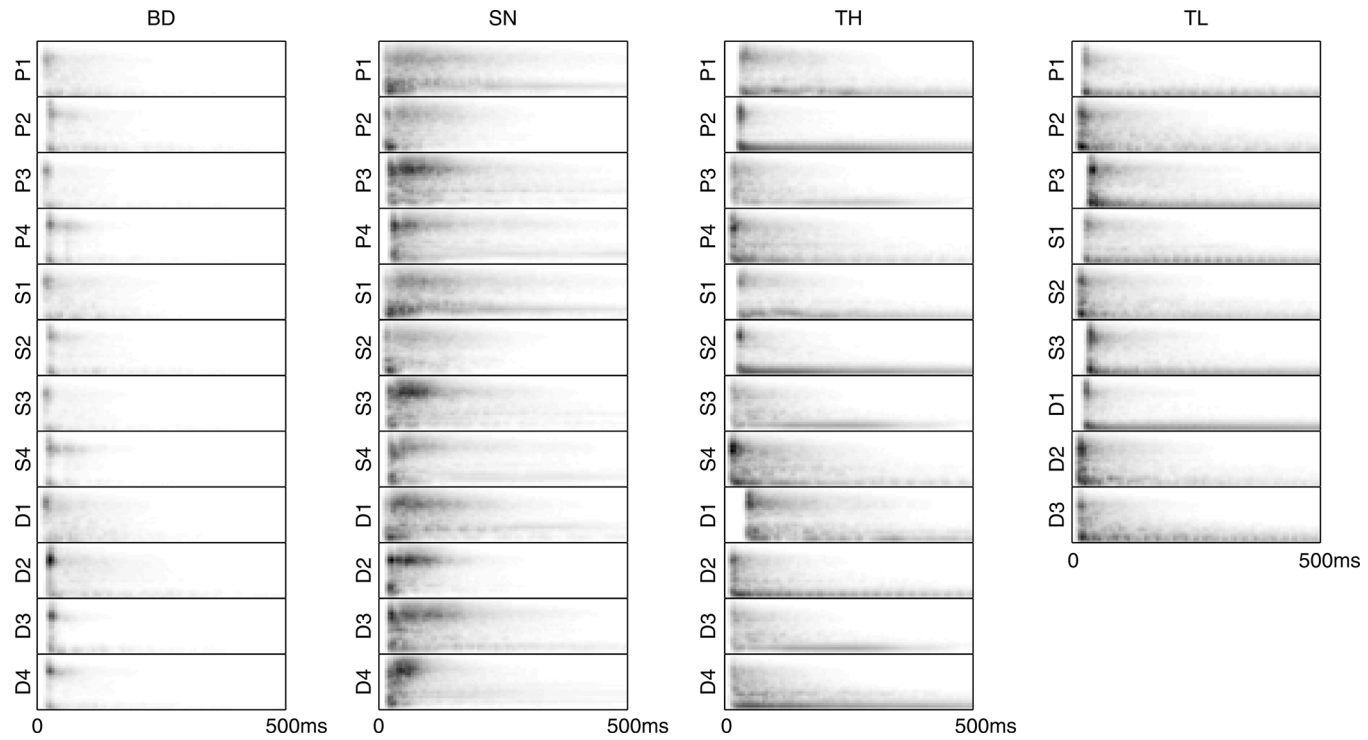

Fig. 1. Auditory images of the stimuli used for the listening test. All samples within a column are from the same instrument class. Black corresponds to high loudness levels, white to silence. Frequency is on the y-axis (in the range of 0–22 kHz). Time is on the x-axis in the range of 0–500 ms. The images were computed (see Section III) using a short-time Fourier transform (STFT) window size of 23 ms and 25% hop size, 24 frequency bands (Bark), loudness in sone, and applying the spectral masking model as well as the outer/middle ear model. The samples marked with P are the manually selected prototypical sounds, those marked with S are the sounds very similar to the respective prototype (e.g., S1 is very similar to P1). The samples marked with D are moderately similar to the respective protoypes.

IV. Ground Truth

To evaluate the models presented in Section V, we use similarity ratings gathered in listening tests and instrument class labels assigned to each sample.

A. Listening Test

For each instrument class, we conducted a listening test where the subjects were asked to rate the similarity of sound pairs. In the following, we describe the participants, the stimuli we used, how we selected the questions we asked, and the implementation of the test (and user interface). At the end of this subsection, we analyze the data.

1) Participants: The 144 (voluntary) participants were mostly colleagues working either at MTG[5] or OFAI.[6] In addition, students from the ESMUC[7] higher music conservatory in Catalonia, and friends were asked to participate. Before the actual listening test, we asked each subject a number of questions regarding previous knowledge which could have an influence on the ratings. Table I summarizes the answers. Several participants took part in more than one test.

2) Stimuli: We selected the stimuli from two commercial drum sample CDs. Most of these samples are very dry, i.e., no additional effects such as reverberation were added to them. We removed those which had effects added to them. Almost all of them are sampled at 44.1 kHz. In a first step, we selected the following number of samples from each instrument class: 76 bass drums (BD), 123 snare drums (SN), 64 high-pitched toms (TH), and 48 low-piched toms (TL). The variances within the same instrument class are due to different diameters (e.g., 20 in and 22 in for bass drums), tuning (high, low), how and where the drums are hit (hard, soft, centered, etc.), the recording environments (small or large ambiance), and different manufacturers (e.g., Noble & Cooley, Yamaha, Premier).

For each of the four instrument classes, we manually selected a small number of prototypical sounds. We selected four (two from each sample CD) for BD, SN, TH, and three (two from the same sample CD) for TL. The reason for selecting one less for TL was that we found less variation in the samples. For each of these prototypical sounds, we selected a very similar sample and one moderately similar sample. Given a prototype P_i, we refer

TABLE I
Number of Participants per Category*

Test	Total (%)	C1 (%)	C2 (%)	C3 (%)	C4 (%)
BD	35 (100)	1 (3)	23 (66)	9 (26)	10 (29)
SN	49 (100)	1 (2)	38 (78)	9 (18)	9 (18)
TH	46 (100)	2 (4)	32 (70)	14 (30)	13 (28)
TL	48 (100)	4 (8)	32 (67)	11 (23)	15 (31)

*The table shows the number of participants per listening test and category. Each participant can be member of none, one, or many categories. The values in brackets are the respective percentage. Category C1 are drummers. C2 are participants with at least a basic level of musical training. C3 participated in previous timbre studies. C4 have experience playing or working with drum samples. Several of the participants took part in more than one test. BD stands for bass drums, SN for snare drums, TH for high pitched toms, and TL for low pitched toms.

[5]Music Technology Group (MTG), Universitat Pompeu Fabra

[6]Austrian Research Institute for Artificial Intelligence (OFAI)

[7]L'Escola Superior de Música de Catalunya (ESMUC)

TABLE II
LENGTH OF THE STIMULI IN MILLISECONDS

	Max	Mean	Min
BD	1632	1147	723
SN	1751	996	386
TH	3374	2228	1280
TL	5987	4248	2005

to the very similar sample as S_i, and we refer to the moderately similar sample as D_i.[8]

In total, we manually selected 12 samples for BD, SN, TH , and nine for TL. The purpose of this subjective selection by the authors is to ensure that the questions (which we ask the participants in the listening test) cover a large range of distances, while minimizing the total number of necessary questions. Alternatively, a larger number of samples could have been selected randomly.

Table II lists the file length statistics of the selected samples. Fig. 1 shows auditory images of the samples. The reason why the samples are so long, and why they do not have the onsets at the same position, is that on one sample CD several samples were given as one long track. We automatically extracted the samples using a very low threshold for the energy. The different onsets in the range of milliseconds do not effect the perception of similarity. However, they do force us to consider alignment techniques described in Subsection III-B9.

3) Questions: We asked the subjects to rate the similarity of the following sample pairs:

A) To measure large differences, we asked the participants to rate the similarity of all pairwise combinations of prototypical samples $P_i - P_j$.
B) To measure large to medium differences, we asked for the similarity of each prototype P_i and the closest D_j sample from a different prototype. (The closest D_j was selected manually.) D_j is generally more similar to P_i than P_i to P_j.
C) To measure medium differences we asked for the similarity of each prototype P_i and its respective moderately similar sample D_i.
D) To measure very small differences, we asked for the similarity of each prototype P_i and its respective very similar sample S_i.

In sum, we asked $18(= 6 + 4 + 4 + 4)$ questions for each of the BD, SN, and TH tests and $12(= 3 + 3 + 3 + 3)$ for the TL test. We selected these pairs with the intention to cover a broad range of differences while minimizing the number of necessary questions. In an ideal case, we would have asked the subjects to rate all possible combinations (for 12 samples that would result in 66 questions). However, we wanted to limit the average time to complete the listening test (including the time to read the instructions) to about 10 min. Furthermore, the task of rating the pairwise similarity of drum samples is rather monotonous, and thus it is difficult to remain concentrated for a longer period of time.

To average out any possible effects presenting the 18 (or 12) questions in a specific order might have, we randomly permutated them for each participant. Furthermore, for each question the order of each sound in the pair was also randomly permutated.

4) Implementation and User Interface: The four different tests were run over a period of more than one year. Each test could be done over the Internet. The participants were asked to use headphones; however, we had no control over the quality of these.

Fig. 2 shows a screenshot of the user interface for the BD test. The interface was implemented as a web-based Java applet. The interface is controlled using keyboard keys "A," "B," and the cursor keys. Given a question with a pair of samples A and B, pressing the respective keyboard character plays the sample. The cursor left and right keys adjust the similarity rating (the position of the slider). The slider has a resolution of nine steps. Using up and down keys the user returns to the previous or proceeds to the next question.

All question pairs (and ratings) are displayed on one screen to allow the subject to easily maintain an overview of the questions. The interface was designed to support users to easily (with only a few keystrokes) jump between different pairs to adjust them if necessary. Most participants needed only very few explanations.

The user interface records the overall time spent before pressing the "Finish" button. Furthermore, for each question the number of times the samples were listened to is recorded. If the user never returns to previous questions to readjust the ratings, there is a dialog box which appears after pressing the "Finish" button which informs the user of this option and asks if the user wants to return to the test and use it.

5) Results: Table III shows the time spent listening to and rating the samples. Completing the ratings in less than one minute is possible for TL because there are only 12 questions. Thus, the subject spent an average time of 3.5 s on each question. Very high values (beyond 30 min) occur because some people started the test, did something else in between, and returned to it later. The participants were told they could use as much time as they like.

Table IV shows statistics of how often each sample was listened to per subject. For all listening tests, the lowest number is 1. The very high numbers are a result of some subjects playing with the interface and experimenting with rhythm patterns. The median values show that half of the users listened to each sample at least three times (if we assume that each of the two samples per question was listened to equally often).

Table V shows how the participants rated the difficulty of the test. The ratings between different tests show little variance with respect to the average. To assess the difficulty, the time spent on the test and the number of times each sample was listened to might be more useful indicators.

Fig. 3 shows the large variance we observed in the subject's ratings (except for very similar pairs where the variance is much lower). Each subject's ratings are normalized such that minimum equals 1 (not similar) and maximum equals 9 (very similar). In the remainder of this paper, we refer to this as the normalized ratings. There are a number of cases where a subject rated the similarity of a pair with 1 while another rated it with 9. As we will discuss later, we believe this is mainly due to the difficulty of the task.

[8]All P_i, S_i pairs were from the same sample CD, and about half of the P_i, D_i pairs came from the same sample CD.

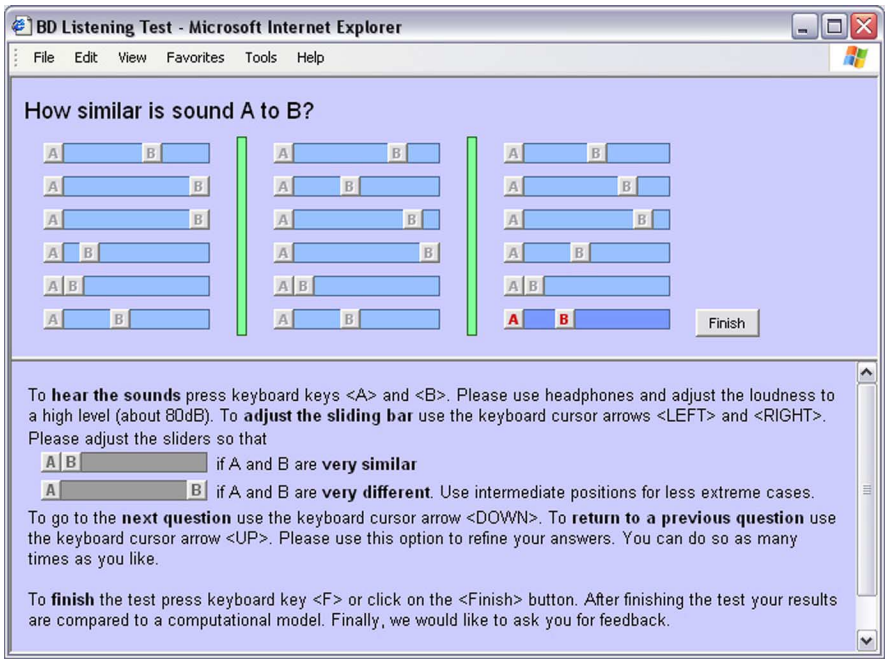

Fig. 2. Screenshot of the user interface used for the BD listening test. The currently selected pair is the last one (lower right). By pressing the "A" and "B" keys, the respective samples are played. Using the curser up key, the user moves back to the question above. The slider is adjusted using the left and right keys. The upper part contains the Java applet, and the lower part contains a summary of the instructions.

TABLE III
TIME SPENT PER TEST*

Test	Min	Median	Mean	Max
BD	2.6	4.8	7.2	37.5
SN	1.6	6.0	9.0	55.1
TH	1.2	3.1	4.9	24.2
TL	0.7	3.1	5.1	61.1

*Time (in minutes) spent listening to samples and rating similarities. The TL test consisted of 12 questions while all others had 18 questions. The median time spent to answer each question is between 10 seconds (for TH) and 20 seconds (for SN).

TABLE IV
NUMBER OF TIMES EACH SAMPLE WAS LISTENED TO PER SUBJECT

Test	Min	Median	Mean	Max
BD	1	6	8.1	152
SN	1	4.5	8.1	178
TH	1	3	4.8	38
TL	1	4	5.6	40

TABLE V
DIFFICULTY OF EACH TEST*

Test	Min	Median	Mean	Max
BD (91.4%)	2	3	3.2	5
SN (95.9%)	1	3	3.0	4
TH (93.8%)	1	3	3.1	4
TL (87.0%)	1	3	3.1	5

*The participants could optionally rate the difficulty after completing the test. The scale ranged from 1 (very easy) to 5 (very hard). The percentages in the first column are the participants who answered the question.

To quantify the consistency of the ratings we compute the average pairwise correlation between the ratings of the subjects (intersubject correlation). This is done for each instrument class. The average correlation between subjects (r_{ss}) is between 0.58 and 0.68 (see Table VI). Because we only asked 18 or 12 questions, many of these intersubject correlations are not significant for BD and TL at the $p < 0.05$ level. An average intersubject correlation of 0.64 for experiments on similarity ratings of synthetic instrument timbres was reported in [12]. Similar intersubject correlations were also reported in [48].

In addition, we compute the mean of the normalized subjects' ratings (MNSR) and measure the average correlation of each subject with the MNSR (r_{ms}). The results are shown in Table VI. We assume that MNSR is the best approximation of the ground truth we have. The average correlation between the subjects and the MNSR is between 0.77 and 0.83 per class. Despite the small number of questions we asked per instrument class, the ratings of almost every subject are significantly correlated with the MNSR. Only in a few cases for TL the correlations are not significant at the $p < 0.05$ level.

6) Variance: Our results show that there is a large variance in the ratings. To better understand the variance, we asked one subject to repeat the test several times.[9] In particular, she repeated the BD test six times over a period of one week. Each time she used the same headphones. She was given no feedback with respect to how her ratings were correlated to her previous ratings or to ratings by other subjects. The mean of the correlations of each individual run with the mean of all six of her normalized runs is 0.82 which is not much higher than the r_{ms} for BD (0.77).

This large variance in her trials indicates that major factors for the variance are independent of the specific headphones or an individually different perception of similarity. A possible factor is the local context, i.e., the order in which the questions are presented. Another factor might be the overall difficulty of the questions (and the resulting difficulty to answer them consistently).

[9]Previously to the listening test the subject did not fit into any of the C1-C4 categories described in Table I.

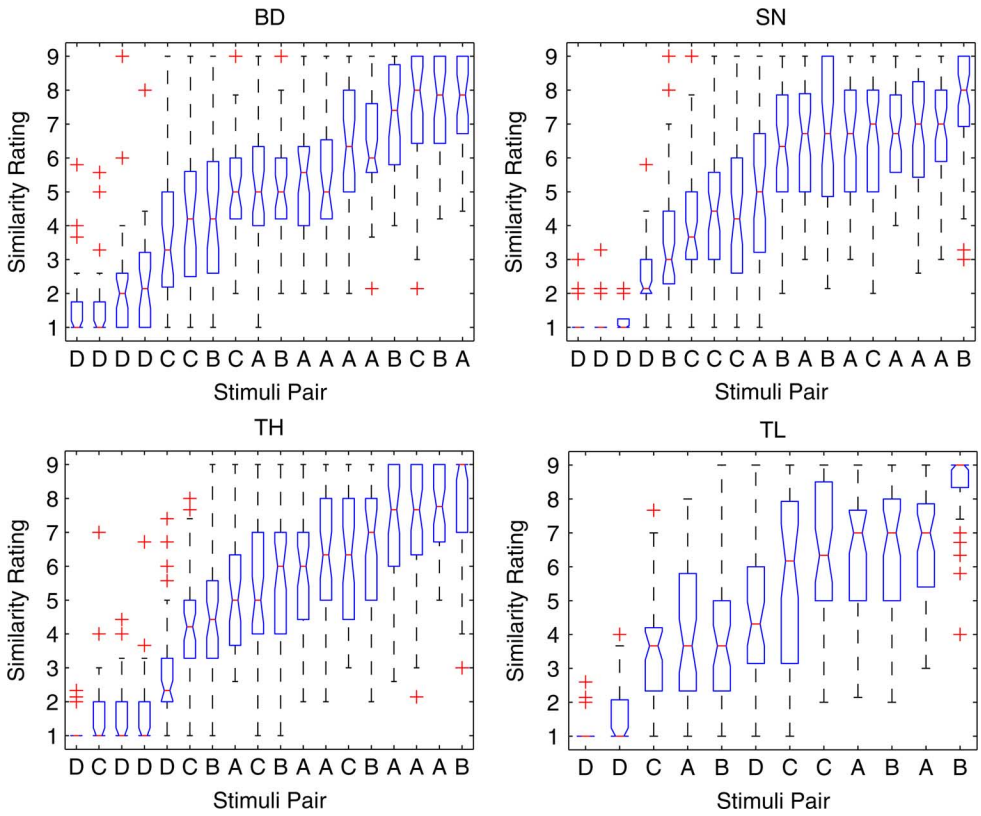

Fig. 3. Boxplots of the normalized similarity ratings. The notches represent a robust estimate of the uncertainty about the medians for box-to-box comparison. Boxes whose notches do not overlap indicate that the medians differ at the 5% significance level. Whiskers extend from the box to the most extreme rating within 1.5 times the interquartile range of the ratings. The stimuli pairs are sorted according to their mean normalized subjects' ratings (MNSR). On the x-axis the pairs are labeled with the four classes described in Subsection IV-A3.

TABLE VI
CORRELATION OF SUBJECT RATINGS WITH THE MNSR AND OTHER SUBJECTS*

	r_{ms}		r_{ss}	
BD	0.77	(0.0027)	0.58	(0.0536)
SN	0.82	(0.0002)	0.67	(0.0116)
TH	0.83	(0.0003)	0.68	(0.0116)
TL	0.78	(0.0075)	0.61	(0.0838)

*Average correlation of the subject ratings with the mean normalized subjects' ratings (r_{ms}) and with the other subject ratings (r_{ss}). The values in the brackets are the average of the respective p-values.

A certain learning effect seems to support that the difficulty is a major factor for the variance. In particular, the correlation of the ratings of each of her trial runs with the MNSR in temporal order of the trials is 0.61, 0.83, 0.75, 0.78, 0.85, and 0.94. Quantifying this learning effect would require further experiments. However, the results indicate that a trained listener might have a higher correlation with the MNSR.

7) Quality of the Ratings: As stated previously, we consider the MNSR to be the best approximation of the ground truth we have. Based on this assumption, we consider ratings which have a higher correlation with the MNSR to be of higher quality. Given the data we have, an important question is how to select participants for future tests to maximize the quality.

Figs. 4 and 5 show the ratings from the subjects having lowest and highest correlations with the MNSR. The highest correlation of a subject with the MNSR (for TH) is 0.97. The lowest correlation (for BD) is 0.48.

Using the information of the questionnaire (see Table I), we analyze if certain groups of subjects perform better than others. Table VII is computed over the four instrument classes. For each category C1–C4 (yes/no), we gather all the correlations. We then compute a two-sample t-test to measure if the means of these samples differ significantly. In particular, we analyze if the members of one group perform better than nonmembers.

In the case of drummers (C1), we cannot claim significant differences because only very few drummers participated. However, the relatively large difference in the correlations indicates that drummers might be the ideal subjects for such tests.

Only in the case of C2 (basic level of musical training) we are able to measure a significant (one-sided p-value $= 0.0187$) difference between subjects that have musical training and subjects that do not. Nonmusically trained subjects tend to have a slightly lower correlation with the MNSR (0.78 compared to 0.81).

In the case of C3 (previous experience in listening tests), the high p-value states that our assumption that subjects with previous experience would perform better is not supported by our data. One possible explanation for this might be that these subjects had different expectations of the test (e.g., in previous tests they might have rated the similarity of a violin and a guitar sound.) and might not have studied the instructions carefully. The difference for C4 (experience playing or working with drum samples) is also smaller than we expected.

8) Discussion: Gathering ground truth data from listening tests is very difficult. As our results show, there is a very high

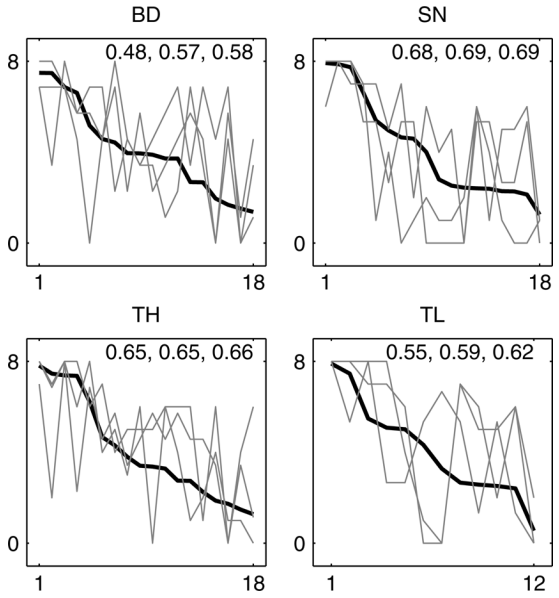

Fig. 4. Normalized ratings of the subjects with the lowest correlation with MNSR. The axes are the same as in Fig. 3 (x-axis are the questions sorted by MNSR, y-axis are the ratings). The thick black line is the average of all subjects (MNSR). Each thin gray line corresponds to one subject. The individual correlations are given in the lower part of the figure.

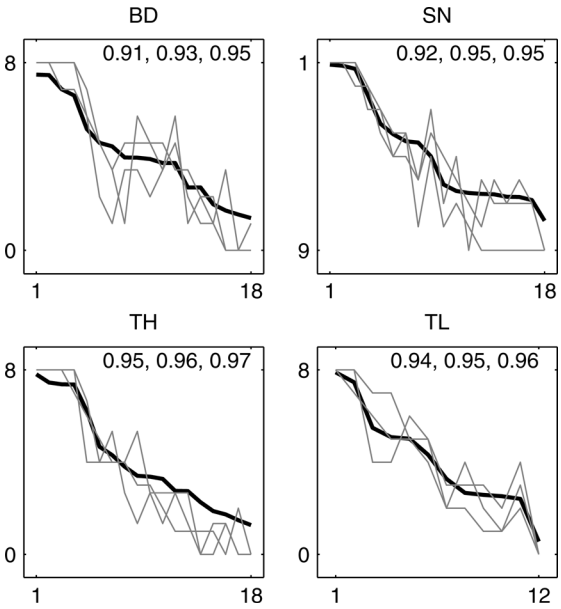

Fig. 5. Same as Fig. 4 for the subjects with the highest correlation with the MNSR.

TABLE VII
CORRELATION OF CATEGORIES OF PARTICIPANTS AND THEIR PERFORMANCE*

	Yes	No	one-sided p-value
C1	0.84	0.80	0.13
C2	0.81	0.78	0.02
C3	0.80	0.81	0.72
C4	0.81	0.80	0.19

*Average correlations (over all 4 tests) for subjects belonging to a certain category (second column) or not (third column) and one-sided p-values for the significance of the difference. The categories (C1–C4) are listed in Table I. The p-values were computed using a two-sample t-test. The sample means were computed from the correlation of each subject's ratings with the MNSR.

variance in the subjects' ratings. Nevertheless, we believe that using the average of all ratings gives us a good approximation of the ground truth. One of the main limitations of our data is that we only have a small number of rated pairs per instrument class.

A number of alternatives exist for designing listening tests to obtain such ground truth. One alternative we implemented is to explicitly define the context by using three samples per question. In particular, given two samples, the subjects were asked to rate (on a scale) to which of these a third sample is more similar to. However, preliminary results from this AB-X tests showed an equally high variance. In addition, the ratings were more difficult to analyze. Alternatively, for example, the subjects could be given all the sounds at once and be asked to organize them by similarity.

B. Instrument Classes

In contrast to the MNSR obtained using listening tests, the instrument class each sample belongs to is very easy to obtain. Furthermore, the instrument class labels are objective. Even if a bass drum sample with a lot of reverberation sounds similar to a snare drum sample, it is a fact that it was generated by a bass drum. However, instrument class data only allows comparisons on the instrument level. Judgments within an instrument class cannot be evaluated directly.

The basic assumption we make, which allows us to use the instrument class labels meaningfully, is that in general the most similar samples to each sample belong to the same class. The same assumption is made, for example, when genre labels are used to evaluate the performance of computational models of similarity for pieces of music (e.g., [49]).

Using instrument class labels transforms the problem into an instrument classification task where a large amount of research has already been published (see Subsection II-C). However, one important constraint is to directly use the similarity computations (e.g., using a nearest neighbor classifier).

The details of how exactly we apply the instrument class ground truth are described in Subsection V-B. Basically, given a large collection of samples and one query sample, we compute a ranked list of the most similar samples from the collection. We then evaluate how many of the top entries in the list are from the same class as the query. Our results show that the resulting evaluation statistics are highly correlated with those we obtain using the MNSR approximation of the ground truth.

1) Data: The samples we use are the same from which we selected the stimuli for the listening test (see Subsection IV-A2). The data originates from two sample CDs. The samples are labeled by the producers. There are a total of 311 samples belonging to four instrument classes: 76 bass drums (BD), 124 snares (SN), 63 toms with a high pitch (TH), and 48 toms with a low pitch (TL).

V. EVALUATION

In this section, we first describe the evaluation based on the data from the listening test, followed by the evaluation using the instrument class data. In Subsection V-C, we discuss our findings.

Overall, we are interested in computational models of similarity for drum sounds which can generate ratings as close as possible to the MNSR. The questions we address are: 1) How

does the MPEG-7 model compare to the auditory image based model? 2) Which impact do the different parameter settings have in computing the auditory images? (What are the optimal parameter settings? What are the differences between the instrument classes?) 3) Can the instrument class data be used to replace listening tests? (How can the instrument class data be used to get the most reliable evaluation statistics?)

A. Listening Test Data

The "ground truth" we use for the evaluations described in this subsection is the MNSR. For each of the classes BD, SN, and TH, we have the similarity ratings for 18 pairs of sounds. For TL, we have 12 rated pairs. In the following, we first describe the evaluation criteria for the models (and different parameter settings). We then present the results for the MPEG-7 model followed by the results for the auditory images. Finally, we analyze if the improvements are statistically significant.

1) Method: Given the MNSR, we are interested in finding models which rate the similarity of the sound pairs the same way. To compare the ratings generated by a model and the MNSR, we compute the correlation coefficient (also known as Pearson's product-moment coefficient). The same approach to evaluate the performance of models for timbre similarity was used, for example, in [12].

To evaluate the impact of the parameters, we use a simple grid search. For each parameter, we define a range of interesting values and evaluate all possible combinations of values for different parameters. An alternative approach to optimize the parameters is to use a gradient search (as used in [13]).

2) Generalization and Overfitting: An important question is how our results can be generalized given that we are primarily focusing on reproducing ratings we obtained in the listening tests. In particular, the question is if the good performance of the auditory image model (which we discuss later on) is too optimistic because the parameters are overfitted to our specific observations, or if they can be generalized.

First of all, it is important to state that we are dealing with very limited degrees of freedom which limit the potential danger of overfitting. The models we investigate have been published previously. The parameter ranges are within reasonable boundaries. In several cases, the parameters we evaluate lead to simplified versions of published models (e.g., using a simple logarithmic scale instead of the Bark scale, or not using any spectral masking models). Nevertheless, we use the following three strategies to obtain a rough estimate of possible overfitting effects.

First, we compare the performance of the model that performs best on average over all four instrument classes with the models individually optimized for each instrument class. Large differences between these differently optimized parameters could indicate overfitting. On the other hand, if the optimal parameters are similar for all classes, then this indicates that the parameters are robust.

Second, we measure performance differences for slightly different parameter settings. Small changes of parameter values should only have a small impact on the performance. If the performance is very good for a large range of parameter values around the optimum, then there is less danger of overfitting. The smoothness of the performance curve with respect to the gradual changes of the parameter values is discussed later on.

TABLE VIII
RESULTS FOR THE MPEG-7 SIMILARLY MODEL*

	w_1	w_2	w_3	BD	SN	TH	TL	AVG
BD	1	6.5	7.5	**0.79**	0.58	0.65	0.78	0.70
SN	0	4	2.5	0.70	**0.84**	0.68	0.78	0.75
TH	0	9.5	7.5	0.74	0.83	**0.69**	0.78	0.76
TL	0	1.5	0	0.35	0.71	0.64	**0.79**	0.62
AVG	0	3	2.5	0.74	0.83	0.69	0.78	**0.76**
ISO	3	6	10	0.75	0.41	0.51	0.76	0.61
LAT	1	0	0	**0.50**	0.28	0.22	0.70	0.43
TC	0	1	0	0.35	**0.71**	**0.64**	**0.79**	**0.62**
SC	0	0	1	0.45	0.57	0.36	0.36	0.43

*Evaluation results for the MPEG-7 similarity model. The columns BD, SN, TH, and TL are the performance in terms of correlation with the MNSR for the respective class. The AVG column is the average correlation. The rows BD, SN, TH, and TL show the correlations when the weights (w_1, w_2, w_3) are optimized for the respective class. ISO marks the row where the weights are set according to the MPEG-7 standard. The last 3 rows show the correlations using only individual dimensions (LAT: log-attack time, TC: temporal centroid, SC: spectral centroid). Bold numbers mark results which are the best within an instrument class (column) and evaluation category (set of rows).

Third, we compare the results of the listening test data-based evaluation with the results from instrument class data-based evaluation. For this second evaluation, the number of samples is much larger, and different evaluation criteria are used. We vary a parameter and compute the evaluation results. The evaluation results for both approaches are highly correlated (see Subsection V-B). This indicates that we have not overfitted the parameters to the data.

Despite these three strategies, our findings cannot be generalized for drum sounds with added effects (such as reverberation) or sounds from other instrument classes because we are not analyzing such sounds.

3) MPEG-7 Results: The results for the MPEG-7 distance model are summarized in Table VIII. The optimal weighting coefficients (w_1, w_2, w_3) for the distance computation were evaluated for each of the four instrument classes. Except for BD the optimal value for w_1 is zero. This indicates that log-attack time is not a useful dimension. However, as shown in the last three rows, for the BD data the LAT by itself yields a higher correlation than SC or TC achieve individually. Overall, when considered individually, the temporal centroid (w_2) is more powerful than the spectral centroid (w_3).

The best combination on average has an average correlation of 0.76. Using the specific weights of the MPEG-7 recommendation yields an average correlation of 0.61. The recommendation works particularly well for BD. In all cases except TH, the best combinations yield correlations which are higher than the average correlation of the subjects' ratings with the MNSR. Given its simplicity, the model works surprisingly well.

In addition, we tried to optimize the MPEG-7 model by implementing variations. One of the best variations we found uses a simple linear combination $(d_{AB} = w_1 d_{\mathrm{LAT}} + w_2 d_{\mathrm{TC}} + w_3 d_{\mathrm{SC}})$. Furthermore, to compute LAT, we set the threshold for t_1 to 80% of the peak energy instead of 100% and computed the spectral centroid using the Bark scale. The highest correlations we found were 0.90 for BD, 0.87 for SN, 0.78 for TH, 0.76 for

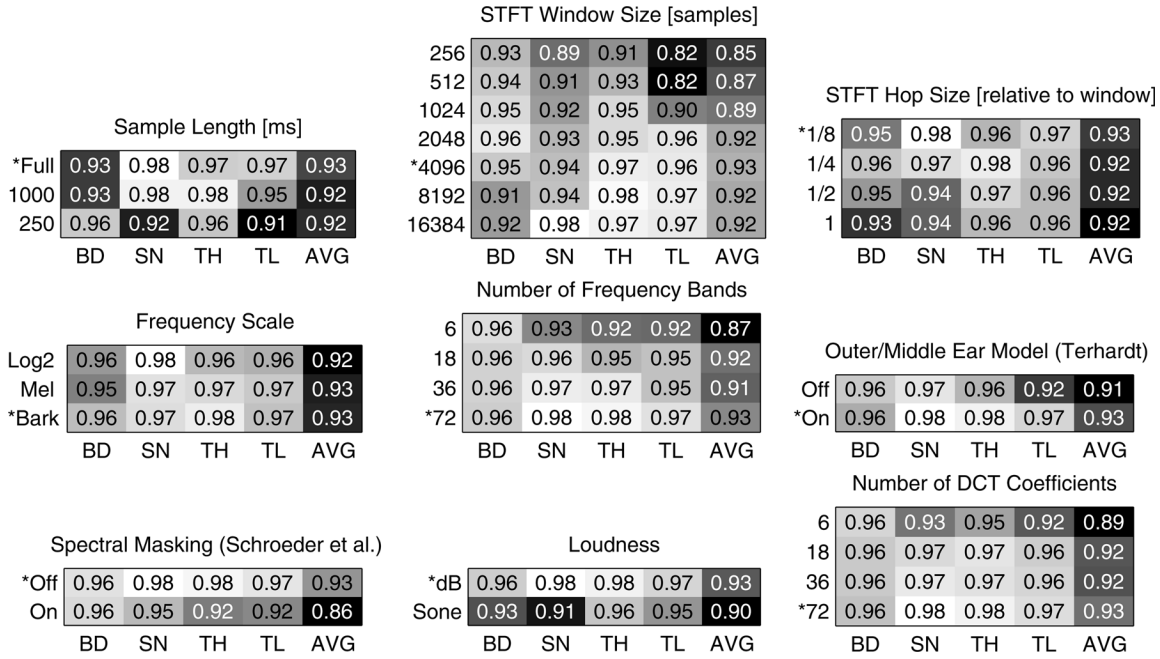

Fig. 6. Evaluation results for the auditory images. Each table summarizes the correlations of 20 160 parameter combinations. For example, in the table in the upper center (STFT window size) the cell in the upper left has the value 0.93. This value is the maximum correlation (of the computed distances with the MNSR) on the BD data that any parameter combination (within the search space) achieves with the window size set to 256 samples. In total, there are 2880 $(= 20160/7)$ different combinations where the STFT window size parameter is set to 256. The settings that produce the best results on average over all instrument classes are marked with a preceding "*."

TL 0.76, and an average of 0.78 for the best combination on average. Although our experiments with variations of the MPEG-7 model were not systematic, this indicates that there is room for further improvements using timbre space based models. However, the results we present in the next subsection show that there is still a large difference compared to the correlations we obtain using auditory images (which yield a average correlation of 0.93 for the best combination on average).

4) Auditory Images Results: In total, we evaluate a parameter space with 20 160 possible combinations. In particular, we evaluate three settings for the sample length, seven for the STFT window size, four for the STFT hop size, three for the frequency scale, four for the number of frequency bands, two for the outer and middle ear model, two for spectral masking, two for the loudness, and four for the number of DCT coefficients. The results, the list of parameters, and the range of values are shown in Fig. 6.

The best parameter settings on average over all instrument classes are the following: full sample length, STFT window size of 4096 samples (93 ms), 1/8th (11.6 ms) STFT hop size, Bark scale, Terhardt's model for the outer and middle ear, no model for spectral masking, loudness in dB, and no DCT compression. The individual instrument class MNSR correlations for this specific model instance are BD: 0.89, SN: 0.94, TH: 0.93, and TL: 0.96.

There are a number of observations. One of them is the positive impact of using only a 250-ms sample length instead of the full samples for BD. (The results show that this effect is not noticeable for other instrument classes.) This might be an indication for the potential of techniques which put a stronger focus on perceptually relevant patterns in the auditory image. The long and hardly audible decay might not be as perceptually relevant as suggested by the sum of energy it contains.

The optimal BD parameters are also different with respect to the frequency resolution. Best results for BD are achieved with a low-frequency resolution (six bands) and a short STFT window size (2048 samples) which leads to less temporal smoothing. Best results for the other instrument classes are achieved with much larger window sizes (up to 16 384 samples) and a high number of frequency bands (72 bands).

Another observation is that using the model by Schroeder *et al.* [45] for spectral masking reduces the correlation, e.g., from 0.98 to 0.92 for TH. In the context of classifying pieces of music according to genre, the negative impact of using this spreading function was also noted in [50].

On the other hand, a parameter which shows very little impact is the frequency scale (logarithmic, Mel, or Bark). The small difference makes the simpler logarithmic model more attractive. Another parameter which shows less impact than expected is the hop size. For the models which perform good on average it seems almost irrelevant if the hop size is 1/8 of the window size, or the whole window size. Only for SN larger differences are noticeable (correlation of 0.98 with 1/8, and correlation of 0.94 with hop size 1/2 or 1).

Figs. 7–9 show the pairwise dependencies between selected parameters. In these figures, each value represents the highest correlation over all combinations where two parameters are set to fixed values.

Fig. 7 shows the influence of the frequency scale on the spectral masking. The spreading function we use was designed for the Bark scale. However, we also apply it to the logarithmic and Mel scale. The reduction in performance is slightly larger

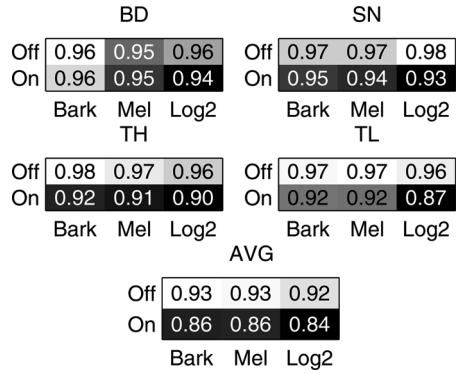

Fig. 7. Evaluation results comparing spectral masking (y-axis) and the frequency scale (x-axis).

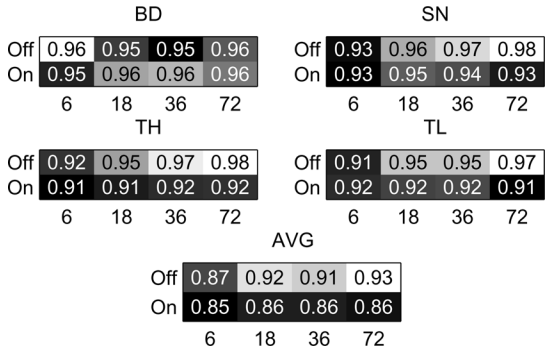

Fig. 8. Evaluation results comparing spectral masking (y-axis) and the number of frequency bands (x-axis).

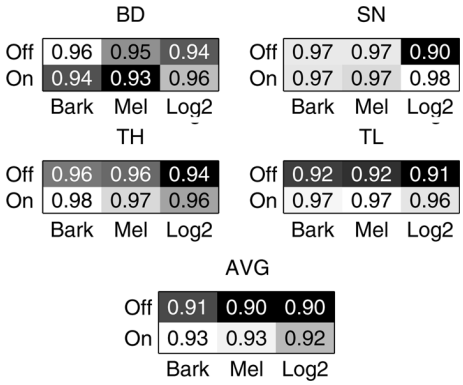

Fig. 9. Evaluation results comparing the outer/middle ear model (y-axis) and the frequency scale (x-axis).

when the spectral masking model is applied in combination with the logarithmic scale. The Bark and Mel scales behave almost identically.

Fig. 8 shows the influence of the number of frequency bands on the spectral masking. When using only few frequency bands, the effects of the spectral masking are negligible, as the limited number of bands already implies strong masking. When increasing the number of frequency bands, the performance remains almost the same when we apply the spectral masking model. The performance increases without the spectral masking model except for BD where spectral masking shows no significant effects.

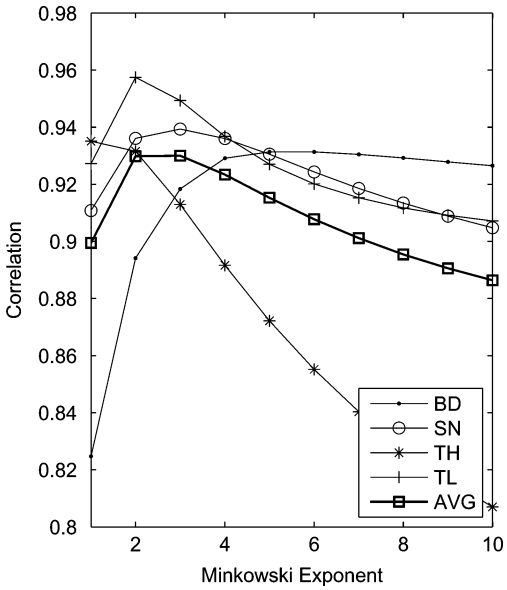

Fig. 10. Correlation for different orders of the Minkowski exponent. The auditory images were computed using the parameter settings which performed best on average (see the second paragraph in Subsection V-A4). Each of the four gray lines shows the performance on one of the instrument classes. The black line is the average performance on all classes.

Fig. 9 shows the influence of the frequency scale on the outer and middle ear model. In most cases (except BD), using the outer and middle ear model improves the performances.

Fig. 10 shows the effect of the exponent p of the Minkowski distance on the correlation of the best on average parameter setting with the MNSR. The exponent p influences the distance computation as follows:

$$d_{\mathrm{AB}} = \left(\sum_{i=1}^{n} |\mathrm{A}_i - \mathrm{B}_i|^p \right)^{1/p} \tag{10}$$

where A and B are two auditory images, and i iterates over all n "pixels" of the images.

The curves show very different characteristics depending on the instrument class. In particular, TH and BD are very different. While TH performs best with an exponent of 1, BD performs best for an exponent of 5 or 6. On average, using the Euclidean distance (an exponent of 2) seems to be a reasonable choice. However, the effects depend very much on the loudness model used. If sone is used instead of dB, then all curves (including BD) are much more similar to the curves of TH and TL, and the peak is on average at 2.

Results which show an optimum exponent of 5 were published in [9]. These findings were confirmed in [13] if an approach is used that does not particularly focus on the onsets. However, using an approach with a special focus on the onsets the observed optimum exponent is 1.

5) Statistical Significance: The results show an increase in average performance over all four instrument classes in terms of correlation with the MNSR of 0.61 (MPEG-7 recommendation) to 0.93 (using auditory images). However, some of the improvements are much smaller, e.g., the highest correlation for the auditory image models optimized specifically for TH is 0.97, while the parameter settings which perform best in average yields 0.96

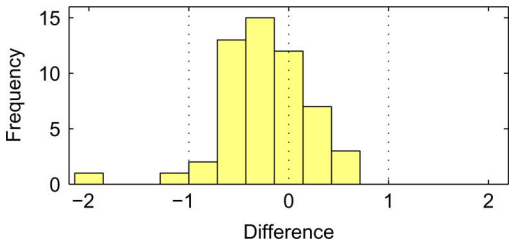

Fig. 11. Histogram of the pairwise difference in errors for the best on average auditory image model and the MPEG-7 recommendation. (Negative values mean that the error of the MPEG-7 recommendation is larger.)

TABLE IX
RESULTS FOR THE AUDITORY IMAGE-BASED SIMILARITY MODEL*

	Average	Individual	one-sided p-value
BD	0.89	0.96	0.008
SN	0.94	0.98	0.039
TH	0.93	0.98	0.026
TL	0.96	0.97	0.133

*Correlations for the best in average and individually optimized parameter combinations for the auditory image based similarity model. The one-sided p-values are computed using a Wilcoxon sign rank test as described in the text.

for TH. Given that we only have a small number of MNSR observations we use statistical tests measure if the improvements are significant.

To measure the significance of the difference we look at the absolute deviation of the model from the MNSR. For each of the $66(=18\times 3+12)$ sound pairs i (from all instrument classes), we have the model rating M_i (e.g., MPEG-7 or auditory image based) and the MNSR (our "ground truth") T_i. We normalize M and T such that the means equal zero and variances equal one. We compute the absolute error of each model (per pair) as $E_i = |M_i - T_i|$.

Given E_i^{α} for the MPEG-7 model and E_i^{β} for the auditory image model which performs best on average we can use a Wilcoxon (sign rank test) to test if the error of the MPEG-7 model is significantly higher. The reason for using the nonparametric Wilcoxon test is that we do not assume that the distribution is necessarily Gaussian.

Fig. 11 shows a histogram of $E_i^{\beta} - E_i^{\alpha}$ over all 66 pairs. That is, the differences in error between the best on average auditory image based model and the MPEG-7 recommendation. The one-sided p-value (median$(E_i^{\beta} - E_i^{\alpha}) < 0$) is 8.2e-6. Thus, the error of the MPEG-7 recommendation is significantly higher on the data we used to optimize the parameters of the auditory image model.

The significance of differences for the different auditory image models is shown in Table IX. The results shows the improvements when using the parameter settings optimized for an individual class compared to the settings which work best on average. The improvements are largest for BD. All improvements except those for TL are significant $(p < 0.05)$.

B. Instrument Class Data

In this subsection, we focus on the question if instrument class information can be used to evaluate the similarity of samples from the same class. The obvious limitation using instrument class data is that it does not allow us to develop instrument-specific similarity models. However, instrument class information is much easier to obtain than similarity ratings from listening tests.

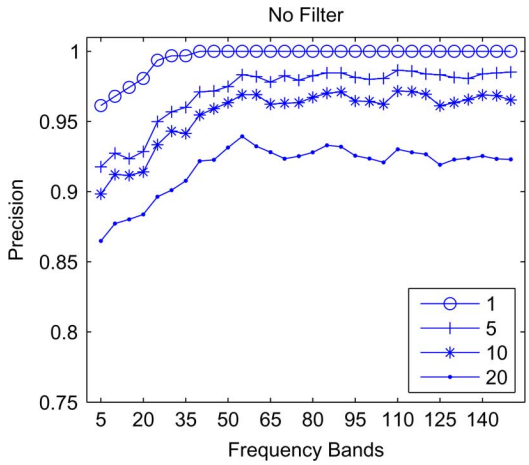

Fig. 12. Precision at $n = 1, 5, 10, 20$ using the instrument class data for auditory images where all parameters are fixed except for the number of frequency bands.

The basic assumption we make is that samples from the same instrument class are on average more similar to each other than to samples from different classes. Similar assumptions are made when genre class information is used to evaluate music similarity (see, e.g., [49] and [51]).

We measure the performance of a model in terms of retrieval precision at n. For each sample, we compute the list of n most similar samples (for $n = 1, 5, 10, 20$). We measure the percentage of n similar samples which are from the same instrument class as the query sample. The precision at n is computed by averaging this count for all 311 samples (the samples are described in Subsection IV-B1).

In addition, we analyze the impact of using a "sample CD filter." In particular, when computing the list of similar samples given a query sample, we ignore all samples from the same sample CD which are from the same instrument class. The motivation for doing this is that samples from the same sample CD might be particularly similar. For example, the only difference between two samples might be that, in one case, the same bass drum was hit "mezzo forte" and "forte" in the other case. This is conceptually very similar to filtering songs from the same artist when using genre class information to evaluate music similarity [49], [51].

To compare the evaluation based on the listening test data with the evaluation based on the instrument class data, we compute the correlation of the respective evaluation statistics for a range of parameters. In particular, we compute the correlation between the precision curves and the MNSR correlation curves while varying the number of frequency bands (for the auditory image-based model) in the range from 5 to 150 in steps of 5. All other parameters are fixed to the settings which perform best on average according to the evaluation using the MNSR (see the second paragraph in Subsection V-A4). A question of particular interest is if the optimal number of frequency bands is the same for both evaluation procedures.

Fig. 12 shows the evaluation results (precision at n) using instrument class information and no sample CD filter. When using more than 40 frequency bands the most similar sample to

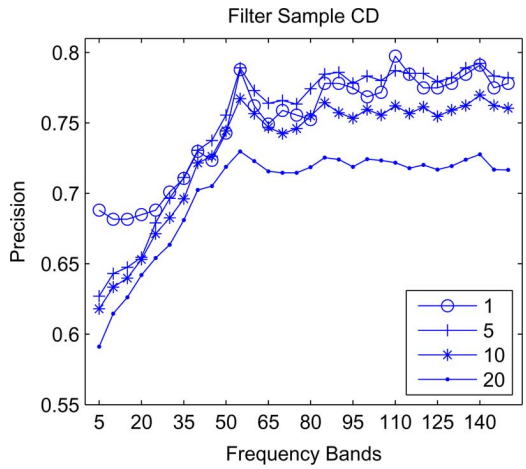

Fig. 13. Precision results using a sample CD filter (compare to Fig. 12).

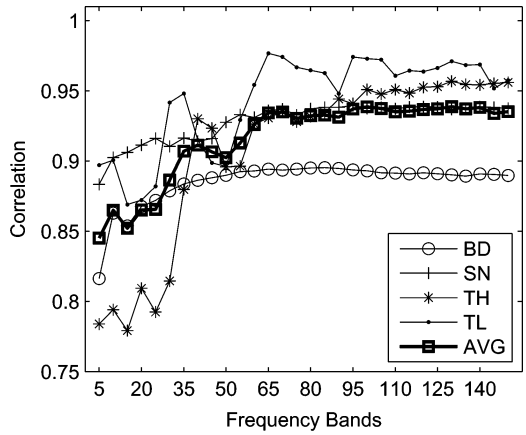

Fig. 14. Correlation of the average normalized subject ratings with the models (where only the number of frequency bands is varied). The gray lines are the correlation for individual instrument classes. The thick black line is the average.

each sample is from the same instrument class. Thus, in terms of instrument classification, using a nearest neighbor classifier would yield 100% accuracy for the classification of the four instrument classes. Overall, the precisions are very high.

Fig. 13 shows the evaluation results (precision at n) using a sample CD filter and instrument class information. The precisions are lower compared to those where no sample CD filter is used (all values are below 80%). One possible explanation is that this is due to the reduced number of relevant samples for each query. Other than the reduced precisions, there are two additional changes. First, the shapes of the performance curves have changed. Second, the relationship between the curves has changed. For example, the differences between the curves for $n = 1$ and $n = 10$ are much smaller.

Fig. 14 shows the correlation of the auditory image-based models and the MNSR. Of particular interest is the average correlation (thick black line). In general, the results show that, for each class, using more frequency bands is better.

To compare the different evaluation procedures, we compute the correlations of their performance curves. In particular, for each of the eight different approaches using instrument class data (four different values for n and using a CD sample filter or not), we compute the correlation with the average curve shown in Fig. 14. The results are given in Table X. All correlations are

TABLE X
CORRELATION OF THE EVALUATION RESULTS*

Sample CD Filter	$n = 1$	$n = 5$	$n = 10$	$n = 20$
Off	0.85	0.96	0.93	0.89
On	0.92	0.96	0.96	0.94

*Correlation of the instrument class based evaluation procedures (precision at n with or without sample CD filter) with the evaluation based on listening test data (correlation with the MNSR).

very significant (all p-values are smaller than 1e-8). In average (over different values of n) the correlations are higher when a sample CD filter is used.

The effects of using a sample CD filter are particularly noticeable for $n = 1$ and $n = 20$. For $n = 5$ there is no difference. However, we do not consider this as an indication that using $n = 5$ will in general produce more reliable results. Instead, we believe that, when using instrument class data to evaluate similarity, it is important to obtain the data from different sample CDs (and thus from different instruments and recording settings) and filter samples from the same CD.

C. Discussion

We have found the following answers to the three questions stated in the beginning of this section: First, the model based on auditory images perform clearly better than the MPEG-7 model which is based on the concept of a timbre space with only few dimensions. The parameters which perform best on average yield a correlation with the MNSR of 0.76 for the MPEG-7 recommendation, and 0.93 for the auditory image-based model.

Second, using auditory images, we have found optimal parameters for each instrument class and parameters which perform best on average. The parameters which work best on average are described in the second paragraph in Subsection V-A4; the parameters which work best for individual instrument classes can be seen in Fig. 6. We have found that in some cases simple approximations can be used to replace more complex models of the auditory system. Furthermore, we have found that the parameter settings optimized for individual instrument classes perform significantly better than the average best parameter settings in all cases except for TL. We found the largest differences between the average best and individual best model for BD. However, given the limited data we have, the average best model seems preferable as it is most likely to generalize well.

Third, we have been able to show that using instrument class data instead of data from listening tests is an attractive alternative to evaluate models of similarity (if the focus is not on developing different models for each instrument class). We have observed correlations of up to 0.96 between the evaluation statistics computed using data from listening tests and statistics computed using instrument class data. Furthermore, our results also show that the instrument class data should be from at least two different sample CDs, and that a "sample CD filter" should be used in the evaluations to produce more reliable results. This confirms earlier findings [29], [52]. However, in contrast to using data from listening tests, using instrument class data does not allow developing instrument class specific models.

1) Generalization and Overfitting: As mentioned in Subsection V-A2, we have been cautious not to claim overoptimistic performances of the auditory image based model. The

following three observations indicate that the performance of the parameters which performs best on average can be generalized. First, Fig. 6 shows that in most cases, slightly different parameter values only have a small impact on the performance. Second, the influence of different parameter values on the performance are similar for most instrument classes (except for some differences between BD and the other instrument classes). Third, using an alternative ground truth and a much larger dataset yields very similar evaluation statistics. However, future studies based on larger scale listening tests are necessary to recommend optimal models for individual instrument classes. Our results show that there is a significant potential for such improvements especially for bass drums.

VI. Conclusion

The primary objective of the work we presented was to evaluate and optimize computational models of similarity for drum samples. Such models are a core technology of retrieval interfaces to large sample libraries [1]. We have studied different models and the impact of parameters. We focused on four percussive instruments which are bass drums, snare drums, high-pitched toms, and low-pitched toms. We did not consider samples to which effects have been added; that is, we only consider very dry samples.

We obtained the ground truth data we used for our evaluations from two sources. Our primary source are listening tests where subjects rated the similarity of pairs of sounds. The secondary source are instrument class labels.

We have found that a similarity model which directly compares aligned auditory images performs significantly better than the MPEG-7 recommendation. We have presented the results of extensive experiments where we have explored a parameter space of 20 160 different possible combinations. We analyzed the impact of using different perceptually motivated computation steps such as using the Bark scale compared to the Mel scale. We found that using a simple approach to approximate effects of spectral masking in the auditory system reduced the overall performance. In addition, our results show little difference between using the simple logarithmic scale to approximate the nonlinear perception of frequency and the more complex Bark and Mel scales. Overall, the model which is most robust in terms of generalization yields an average correlation of 0.93 with human similarity ratings. We believe that this is high enough for the applications we are targeting.

Furthermore, we showed that instrument class data is an attractive alternative source for ground truth data compared to data gathered in listening tests. Our results show correlations of up to 0.96 for the evaluation statistics.

There are various directions for future work. We believe a particularly interesting direction is to investigate how focusing on the onsets (or other perceptually relevant parts of the signal) can increase the performance. This could be done, for example, using techniques suggested in [13]. In general, focusing on specific characteristics which attract the listeners attention seems a viable direction for future work. An alternative direction could be to improve the simple alignment approach we have used. For example, dynamic time warping techniques could be used instead. Furthermore, an interesting direction seems to be to investigate models which allow higher frequency resolutions for lower frequencies and at the same time higher temporal resolution for higher frequencies. Additionally, it would be interesting to develop models which can deal with drum sounds to which effects have been added (such as reverberation). Finally, these models should be connected to generic models of categorization and similarity that have been proposed in cognitive science (e.g., [53]).

Acknowledgment

The authors would like to thank the participants of the listening test, most of which work or study either at MTG, OFAI, or ESMUC. Furthermore, the authors wish to thank the anonymous reviewers for very helpful comments.

References

[1] E. Pampalk, P. Hlavac, and P. Herrera, "Hierarchical organization and visualization of drum sample libraries," in *Proc. 7th Int. Conf. Digital Audio Effects*, Naples, Italy, 2004, pp. 378–383.

[2] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, ser. Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage Publications, 1978, pp. 7–11.

[3] J. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1270–1277, 1977.

[4] S. McAdams, S. Winsberg, S. Donnadieu, G. D. Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychol. Res.*, vol. 58, pp. 177–192, 1995.

[5] N. R. Misdariis, B. K. Smith, D. Pressnitzer, P. Susini, and S. McAdams, "Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres," in *Proc. 16th Int. Congr. Acoust.*, Woodbury, NY, 1998, pp. 3005–3006.

[6] S. Lakatos, "A common perceptual space for harmonic and percussive timbre," *Percept. Psychophys.*, vol. 62, no. 7, pp. 1426–1439, 2000.

[7] J. W. Beauchamp and S. Lakatos, "New spectro-temporal measures of musical instrument sounds used for a study of timbral similarity of rise-time and centroid-normalized musical sounds," in *Proc. 7th Int. Conf. Music Percept. Cognition*, Sydney, Australia, 2002, pp. 592–595.

[8] B. Feiten, R. Frank, and T. Ungvary, "Organisation of sounds with neural nets," in *Proc. Int. Comput. Music Conf.*., San Francisco, CA, 1991, pp. 441–444, ICMA.

[9] B. Feiten and S. Günzel, "Distance measure for the organization of sounds," *Acustica*, vol. 78, no. 3, pp. 181–184, 1993, (research notes).

[10] B. Feiten and S. Günzel, "Automatic indexing of a sound database using self-organizing neural nets," *Comput. Music J.*, vol. 18, no. 3, pp. 53–65, 1994.

[11] P. Cosi, G. De Poli, and G. Lauzzana, "Auditory modeling and self-organizing neural networks for timbre classification," *J. New Music Res.*, vol. 23, pp. 71–98, 1994.

[12] P. Toiviainen, M. Kaipainen, and J. Louhivuori, "Musical timbre: Similarity ratings correlate with computational feature space distances," *J. New Music Res.*, vol. 24, no. 1, pp. 282–298, 1995.

[13] P. Toiviainen, "Optimizing auditory images and distance metrics for self-organizing maps," *J. New Music Res.*, vol. 25, no. 1, pp. 1–30, 1996.

[14] G. De Poli and P. Prandoni, "Sonological models for timbre characterization," *J. New Music Res.*, vol. 26, no. 2, pp. 170–197, 1997.

[15] R. D. Patterson, "Auditory images: How complex sounds are represented in the auditory system," *J. Acoust. Soc. Jpn.*, vol. 21, no. 4, pp. 183–190, 2000.

[16] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ—The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–27, 2000.

[17] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, May 1992.

[18] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 40, no. 12, pp. 963–978, 1992.

[19] E. Skovenborg and S. H. Nielsen, "Evaluation of different loudness models with music and speech material," in *Proc. 117th Audio Eng. Soc. Conv.*, San Francisco, CA, 2004, preprint 6234.

[20] B. Moore, B. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, 1997.

[21] P. Herrera, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instruments," in *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.
[22] D. FitzGerald and J. Paulus, "Unpitched percussion transcription," in *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.
[23] E. Wold, T. Blum, D. Kreislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.
[24] K. Martin, "Sound-source recognition: A theory and computational model," Ph.D. dissertation, Mass. Inst. Technol., , MA, 1999.
[25] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1933–1941, 1999.
[26] J. C. Brown, O. Houix, and S. McAdams, "Feature dependance in the automatic identification of musical woodwind instruments," *J. Acoust. Soc. Amer.*, vol. 109, no. 3, pp. 1064–1072, 2001.
[27] F. Gouyon, F. Pachet, and O. Delerue, "On the use of zero-crossing rate for an application of classification of percussive sounds," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Verona, Italy, 2000 [Online]. Available: http://www.iua.upf.es/mtg/publications/dafx00-gouyon.pdf
[28] P. Herrera, A. Yeterian, and F. Gouyon, "Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques," in *Proc. Int. Conf. Music Artif. Intell. (ICMAI).*, Edinburgh, U.K., 2002, pp. 69–80.
[29] P. Herrera, A. Dehamel, and F. Gouyon, "Automatic labeling of unpitched percussion sounds," in *Proc. Audio Eng. Soc., 114th Convention*, Amsterdam, The Netherlands, 2003 [Online]. Available: http://www.iua.upf.es/mtg/publications/AES114-Herrera2003.PDF
[30] M. A. Loureiro, H. B. de Paula, and H. C. Yehia, "Timbre classification of a single musical instrument," in *Proc. ISMIR Int. Conf. Music Inf. Retrieval*, Barcelona, Spain, 2004, pp. 546–549.
[31] D. Bolger and N. Griffith, "Multidimensional timbre analysis of shakuhachi honkyoku," in *Proc. Conf. Interdisciplinary Musicol. (CIM05)*, Montréal, QC, Canada, Mar. 10–12, 2005.
[32] R. A. FitzGerald and A. T. Lindsay, "Tying semantic labels to computational descriptors of similar timbres," in *Proc. Sound Music Comput. (SMC05)*, Paris, France, Oct. 2004.
[33] A. Tindale, "Classification of snare drum sounds using neural networks," Master's thesis, Faculty of Music, Music Technology Department, McGill University, Montréal, QC, Canada, Sep. 2004.
[34] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7," in *Proc. Int. Comput. Music Conf. (ICMC)*, Berlin, Germany, 2000 [Online]. Available: http://www.iua.upf.es/mtg/publications/icmc00-perfe.pdf
[35] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. 245–248.
[36] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrogram: A new musical instrument recognition technique without using onset detection nor f0 estimation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, May 2006, pp. V-229–V-232.
[37] K. Yoshii, M. Goto, and H. G. Okuno, "Automatic drum sound description for real-world music using template adaptation and matching methods," in *Proc. ISMIR Int. Conf. Music Inf. Retrieval*, Barcelona, Spain, 2004, pp. 184–191.
[38] K. Yoshii, M. Goto, and H. G. Okuno, "INTER:D: A drum sound equalizer for controlling volume and timbre of drums," in *Proc. 2nd Eur. Workshop Integration Knowledge Semantic Digital Media Technol.*, London, U.K., Nov. 2005, pp. 205–212.
[39] M. Slaney, "Semantic-audio retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 4108–4111.
[40] E. Terhardt, "Calculating virtual pitch," *Hear. Res.*, vol. 1, pp. 155–182, 1979.
[41] A. de Cheveigné, "Pitch perception models," in *Pitch*. New York: Springer-Verlag, 2004.
[42] S. S. Stevens, J. Volkman, and E. B. Newman, "A scale for the measurement of the psychological magnitude of pitch," *J. Acoust. Soc. Amer.*, vol. 8, pp. 185–190, 1937.
[43] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 2nd ed. Berlin: Springer, 1999.
[44] E. Pampalk, "A Matlab toolbox to compute music similarity from audio," in *Proc. ISMIR Int. Conf. Music Inf. Retrieval*, Barcelona, Spain, Oct. 10–14, 2004, pp. 254–257.
[45] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1647–1652, 1979.
[46] E. Ambikairajah, A. Davis, and W. Wong, "Auditory masking and mpeg-1 audio compression," *Electron. Commun. Eng. J.*, vol. 9, no. 4, pp. 165–175, Aug. 1997.
[47] R. Bladon and B. Lindblom, "Modeling the judgment of vowel quality differences," *J. Acoust. Soc. Amer.*, vol. 69, no. 5, pp. 1414–1422, 1981.
[48] P. Iverson and C. Krumhansl, "Isolating the dynamic attributes of musical timbre," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 2595–2603, 1993.
[49] E. Pampalk, "Computational models of music similarity and their application in music information retrieval," Ph.D. dissertation, Vienna Univ. Technol., Vienna, Austria, Mar. 2006.
[50] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification," in *Proc. ISMIR Int. Conf. Music Inf. Retrieval*, London, U.K., 2005, pp. 34–41.
[51] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proc. ISMIR Int. Conf. Music Inf. Retrieval*, London, U.K., 2005, pp. 628–633.
[52] A. Livshin and X. Rodet, "The importance of cross database evaluation in sound classification," in *Proc. ISMIR Int. Conf. Music Inf. Retrieval*, Baltimore, MD, 2003, pp. 241–242.
[53] R. L. Goldstone, "The role of similarity in categorization: Providing a groundwork," *Cognition*, vol. 52, pp. 152–157, 1994.

Elias Pampalk received the Doctor of Computer Science degree from the Vienna University of Technology, Vienna, Austria, in March 2006.

Until March 2007, he was a PostDoctoral Research Scientist at the National Institute of Advanced Industrial Science and Technology (AIST). Currently, he is with Last.fm, Ltd., London, U.K.

Perfecto Herrera received the degree in psychology from the University of Barcelona, Barcelona, Spain, in 1987, where he is pursuing the Ph.D. degree. His studies have focused on computer music, sound engineering, and audio postproduction.

He was with the University of Barcelona as a Software Developer and an Assistant Professor. He has been working in the Music Technology Group, University of Pompeu Fabra, Barcelona, since its inception in 1996, first as the person responsible for the sound laboratory/studio, then as a Researcher. He worked in the MPEG-7 standardization initiative from 1999 to 2001. Then, he collaborated in the EU-IST-funded CUIDADO project, contributing to the research and development of tools for indexing and retrieving music and sound collections. This work was somehow continued and expanded as Scientific Coordinator for the Semantic Interaction with Music Audio Contents (SIMAC) project, again funded by the EU-IST. He is currently the Head of the Department of Sonology, Higher Music School of Catalonia (ESMUC), where he teaches music technology and psychoacoustics. His main research interests are music content processing, classification, and music perception and cognition.

Masataka Goto received the Doctor of Engineering degree from Waseda University, Tokyo, Japan, in 1998.

He then joined the Electrotechnical Laboratory (ETL), which was reorganized as the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, in 2001, where he has been a Senior Research Scientist since 2005. He served concurrently as a Researcher in Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation (JST), from 2000 to 2003, and has been an Associate Professor of the Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba, since 2005.

Dr. Goto has received 20 awards, including the IPSJ Best Paper Award, IPSJ Yamashita SIG Research Awards, and Interaction 2003 Best Paper Award.