

Constructing high-level perceptual audio descriptors for textural sounds

Thomas Grill

Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6, A-1010 Vienna/Austria
thomas.grill@ofai.at

ABSTRACT

This paper describes the construction of computable audio descriptors capable of modeling relevant high-level perceptual qualities of textural sounds. These qualities – all metaphoric bipolar and continuous constructs – have been identified in previous research: *high–low*, *ordered–chaotic*, *smooth–coarse*, *tonal–noisy*, and *homogeneous–heterogeneous*, covering timbral, temporal and structural properties of sound. We detail the construction of the descriptors and demonstrate the effects of tuning with respect to individual accuracy or mutual independence. The descriptors are evaluated on a corpus of 100 textural sounds against respective measures of human perception that have been retrieved by use of an online survey. Potential future use of perceptual audio descriptors in music creation is illustrated by a prototypic sound browser application.

1. INTRODUCTION

For music-making in the digital age, techniques for efficient navigation in the vast universe of digitally stored sounds have become indispensable. These imply appropriate characterization, organization and visual representation of entire sound libraries and their individual elements. Widely used strategies of sound library organization include semantic tagging or various techniques from the field of Music Information Retrieval (MIR) to automatically *classify* and *cluster* sounds according to certain *audio descriptors* which characterize the signal content. Especially interesting are descriptors that are aligned with human perception, since they enable immediate comprehension without the necessity of translation or learning. Such perceptual descriptors are also interesting for applications in musical creation with digital sounds, where intuitive access to certain sound characteristics may be desired.

In this paper, we will detail the construction of computable descriptors capable of modeling relevant perceptual qualities of sound. We will restrict our focus to textural sounds, that is, sounds that appear as stationary (in a statistical sense), as opposed to evolving over time. This broad class of sounds of diverse natural or technical origin (cf. [1]) is interesting for electroacoustic composers, sound designers or electronic music performers because of

its neutrality and malleability, functioning as source material for many forms of structural processing.

The structure of the paper is as follows: In the next Section 2 we will contextualize our research endeavor and describe the fundamentals of our approach. This is followed by a detailed description of our methods (Section 3) and a thorough evaluation of our experimental results in Section 4. Section 5 presents a prototypic application of our research. Finally, Section 6 concludes with a summary of the findings and possible implications for the future.

2. CONTEXT

2.1 Perceptual qualities of textural sounds

For the following, we refer to recent research [2] of our group. We have elicited a number of so called *personal constructs* that are relevant to human listeners for the description and distinction of textural sounds. More precisely, those constructs are *group norms* that are shared by the range of persons – all trained listeners in that case – who have participated in the experiments. The most significant constructs we found are listed in Table 1, sorted by the degree of agreement among listeners. As can be seen, each of the constructs is bipolar, spanning a continuous range from one extreme to the other. The construct *natural–artificial* is somewhat special as it does not refer to an objective, measurable quality of sound, but rather to the source of its production. Since we are interested in automatically computable quantities we can not consider this construct for the present paper. The listed qualities describe spectral/timbral (*high–low*, *tonal*¹–*noisy*) and structural/temporal (*ordered–chaotic*, *smooth–coarse*, *homogeneous–heterogeneous*) aspects of sound. Figure 1 shows a correlation matrix of those perceptual constructs which reveals that the qualities are not fully independent. Substantial off-diagonal values for the correlations of *ordered–chaotic/homogeneous–heterogeneous* and *tonal–noisy/smooth–coarse* result from some degree of similarity as perceived by the listeners in the experiments.

2.2 Audio descriptors

Within the domain of MIR a large selection of audio descriptors is readily available (see [3, 4]). However, we found in [2] that only the constructs *high–low* and *smooth–coarse* show significant correlations of above 0.6 with some of the low-level audio descriptors operating

¹ *Tonal*, as in *tonal language* is synonymous to *pitched*

Construct	Agreement α
high–low	0.588
bright–dull	
ordered–chaotic	0.556
coherent–erratic	
natural–artificial	0.551
analog–digital	
smooth–coarse	0.527
soft–raspy	
tonal–noisy	0.523
homogeneous–heterogeneous	0.519
uniform–differentiated	

Table 1: Perceptual qualities (bipolar *personal constructs*) with their synonymous alternatives as identified in [2]. Constructs are ordered by decreasing agreement (Krippendorff’s α) – top ones have been rated more consistently by the subjects than those at the bottom.

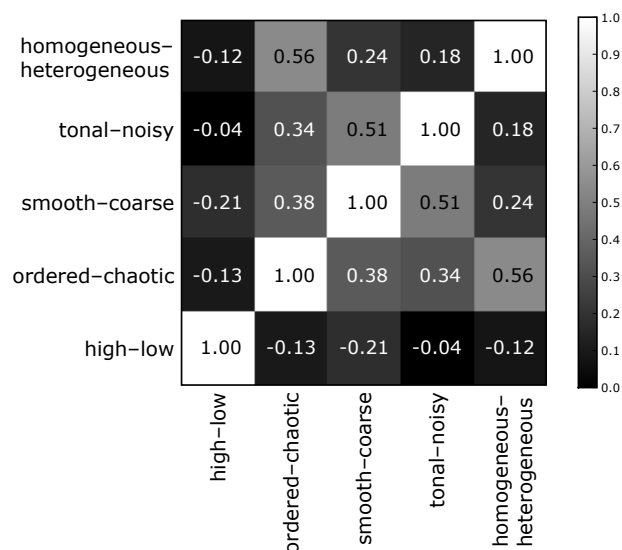


Figure 1: Pearson correlations between perceptual qualities. The smallest significant correlation value (at $\alpha = 0.05$, two-tailed) is ± 0.049 . Taken from [2].

on spectral properties. The other constructs, especially the structural ones (*ordered–chaotic* and *homogeneous–heterogeneous*) can not be represented satisfactorily by available low-level descriptors. A good account of other timbral, but also rhythmic and pitch content features is given in [5, 6].

The overwhelming bulk of the literature dealing with audio features is concerned with ‘song’ characterization and classification (cf. [7]), where two of the predominant tasks are genre classification [8, 9] or emotion resp. *mood* classification [10–12]. They are usually tackled using a set of audio descriptors combined with machine learning algorithms (like Support Vector Machines and others, see [13, 14]) to unearth potential relationships between feature combinations and the target classes. In most cases the classification is performed on discrete classes, either genre classes like ‘rock’, ‘pop’, ‘jazz’, ‘classical music’, ‘world music’, etc., or mood classes like ‘happy’, ‘sad’,

‘dramatic’, ‘mysterious’, ‘passionate’, and others.

While the genre concept does not seem to be applicable to general – especially textural – sound, the task of modeling emotion resp. *affect* in sound and music is somewhat comparable to the task of modeling perceptual qualities. With the existence of a continuous representation of emotion in the *valence–arousal* plane [15, 16], [17] formulate music emotion recognition (MER) as a regression problem to predict such arousal and valence values. They test both linear regression and Support Vector Regression (SVR, see [18]) based on a selection of 18 – mostly spectral – musical features. Alternative approaches have been published by [19], [20], and [21]. In preliminary experiments we tried to model our perceptual qualities using SVR techniques with a range of low-level descriptors (especially those by [6]) but could not achieve significant correlations to human ratings – most probably because the desired high-level characteristics are not represented by the employed low-level descriptors.

An alternative strategy would be to construct specialized descriptors by engineering algorithms to fit to given experimental data [22]. However, this task becomes considerably easier if the underlying mechanisms are known and understood, in which case one can try to model algorithms using expert knowledge in a more or less well-directed manner.

3. METHOD

The general idea of this paper is to construct perceptual descriptors from a uniform underlying representation for the digital audio data with a few processing steps and a small number (to avoid the danger of over-fitting from the start) of adjustable parameters. These parameters we can tune to match perceptual ratings of sounds from a representative corpus.

3.1 Sound corpus and perceptual ground-truth

The elements of the sound corpus have been taken from a large collection of mostly environmental and abstract sounds used for electroacoustic music composition and performance of the author. A total of 100 sounds² have been selected from the library, fulfilling the criteria of being textural and not strongly exhibiting their provenience. The rationale for the latter is that a highly evident origin of sound production (e.g. recognizable materials, cultural or natural contexts etc.) could distract listeners from *objective* qualities inherent in the sound matter. This is strongly related to an *acousmatic* ‘reduced listening’ mode as formulated by Pierre Schaffer [23]. The sounds have been normalized in regard to perceived loudness and mixed down to mono.

Each sound has been rated with regard to the set of perceptual constructs. For this, an online survey³ has been conducted (see [2] for details), widely disseminated among scholars and artists in the domain of sound creation and analysis. On average, each sound has been rated 16 times.

² <http://grrrr.org/data/research/texvis/texqual.html>, retrieved 2012-04-11.

³ Still online at <http://grrrr.org/test/classify>, retrieved 2012-04-11.

We employed a normalization per user to unify mean and variance of the ratings, and used a 0.75-quantile (centered at the mean) per sound to eliminate outliers. From that we calculated mean and standard deviation per quality and sound which we used for the evaluation in Section 4 below.

3.2 Underlying time-frequency representation

We build on a time-frequency (spectrogram) representation, using a variant of the Constant-Q Non-stationary Gabor transform (NSGT) [24]⁴, in our case employing a perceptually motivated Mel frequency scale. By carefully analyzing each of the target constructs for respective spectral and temporal characteristics imprinted into the spectrograms of various sounds, step-by-step we construct techniques to measure the individual qualities. In order to arrive at – potentially non-linear – scales that coincide with listeners’ perception, at several points we introduce scale warping exponents along axes of time, frequency, loudness etc. For that, we will often use the *generalized mean*

$$M_i^p(x_i) = \left(\frac{1}{n} \sum_i x_i^p \right)^{1/p} \quad (1)$$

where a small value p compresses and, conversely, large p expands the magnitude range of the summed coefficients. On the other hand, M^p with large negative p provides us with an equivalent to min, and with large positive p with max, while being continuously differentiable, which is a prerequisite for many optimization schemes. The notation for the running index i below the operator is used to indicate the axis of operation in case the operands are multi-dimensional.

From monophonic digital audio data s we calculate a power spectrogram, with $f_{min} = 50$ Hz (77.8 Mel), $f_{max} = 15$ kHz (3505 Mel) and a frequency resolution of 100 bins ($\Delta f = 34.3$ Mel) in this range. The temporal resolution is set to 10 ms. We also employ psychoacoustic processing using an *outer-ear transfer function* and calculation of *perceived loudness* in Sone units⁵.

$$c_{t,f} = \text{psy} \left(|\text{NSGT}(s)|^2 \right), f = [f_{min}, f_{max}] \quad (2)$$

We normalize the sonogram by the mean (i.e. M^1) value of all time-frequency coefficients in the sonogram.

$$\bar{c}_{t,f} = \frac{c_{t,f}}{M_{t',f'}^1(c_{t',f'})} \quad (3)$$

We also allow attenuation of the spectrum (e.g. for a bass or treble boost) by a small number of interpolated factors Φ equally spaced on the Mel scale.

$$\tilde{c}_{t,f} = \bar{c}_{t,f} \text{ att}_{\Phi}(f) \quad (4)$$

Currently, we use only two factors, one given for the low end and one for the high end of the frequency scale allowing merely a tilt of bass vs. treble frequencies. These two

⁴ implemented in the Python programming language – refer to <http://grrrr.org/nsgt>, last checked 2012-05-21.

⁵ analogous to http://www.pampalk.at/ma/documentation.html#ma_sone (retrieved 2012-04-07), with full scale at 96 dB SPL, omitting spectral spread.

attenuation factors are interpolated over the Mel spectrum in logarithmic magnitudes.

$$\text{att}_{\Phi}(f) = \exp \left(\log \Phi_{low} + \frac{f - f_{min}}{f_{max} - f_{min}} (\log \Phi_{high} - \log \Phi_{low}) \right) \quad (5)$$

The filter coefficients Φ will be tuned specifically for each of the following perceptual features.

3.3 Descriptor for *high–low*

As we have found in [2], this audio feature is quite well represented by the existing *PerceptualSharpness* audio descriptor, which is the “perceptual equivalent to the spectral centroid but computed using the specific loudness of the Bark bands” [4]. The latter is already provided by our chosen time-frequency representation, so we only have to calculate the spectral centroids, but not without applying some tunable warping coefficients.

First we scale the loudness range of the time-frequency components by applying a power ξ .

$$\check{c}_{t,f} = \hat{c}_{t,f}^{\xi} \quad (6)$$

Then we calculate the centroid for each time frame with also warping the frequency axis by applying a power η .

$$\mu_t = \frac{\sum_f f^{\eta} \check{c}_{t,f}}{\sum_f \check{c}_{t,f}} \quad (7)$$

Finally, we take the negative mean of all centroids – negative, as high centroid values correspond to the lower end of the *high–low* continuum:

$$D_{high-low} = -M_t^1(\mu_t) \quad (8)$$

3.4 Descriptor for *ordered–chaotic*

We suspect that the perception of order vs. chaos is not sensitive with regard to intensity, but rather to temporal structure (cf. [25]). Hence, we first *detrend* the sonogram with respect to loudness by high-pass filtering along the time axis. This is done by subtracting a convolution of the mean loudness (over all frequencies) with a Gaussian kernel \mathcal{G} of half-width σ .

$$\check{c}_{t,f} = \hat{c}_{t,f} - \left(M_f^1(\hat{c}_{t,f}) \star \mathcal{G}_{\sigma}(t) \right) \quad (9)$$

Then we slice the resulting sonogram into pieces s_i of uniform duration n with a *hop size* of m by use of a rectangular (boxcar) window function w_n , allowing an optional temporal offset ν in the range $[0, \nu_{max}]$:

$$s_{i,\nu,t,f} = \sum_{t'} w_n(t' - (im + \nu)) \check{c}_{t',f} \quad (10)$$

We use this offset ν to shift sonogram slices s_i against each other in time and calculate a *distance function*, with minima indicating repetitions of time-frequency structures. This is done by use of a generalized mean with an exponent

$\xi < 0$ over all slice elements, specifically exposing small differences.

$$\delta_{i,\nu} = M_{t,f}^{\xi} |s_{i,\nu,t,f} - s_{i,0,t,f}| \quad (11)$$

This yields a set of distance curves as a function of the time shift ν , one for each sonogram slice s_i . For ‘‘ordered sound’’ they are expected to show minima at the same shift positions over all the traces δ_i . We account for evolutions in time by not matching the whole set of curves, but rather comparing only consecutive instances. To enforce magnitude invariance of the curves δ_i we subtract the means

$$\bar{\delta}_{i,\nu} = \delta_{i,\nu} - M_{\nu}^1(\delta_{i,\nu}) \quad (12)$$

before calculating mean distances using an exponent η :

$$\gamma_i = M_{\nu}^{\eta} |\bar{\delta}_{i+1,\nu} - \bar{\delta}_{i,\nu}| \quad (13)$$

We will weight these similarity measures by the mean magnitude of the minima of the curves (calculated by a generalized mean with a large magnitude negative exponent α) to factor in the amount of repetitive similarity,

$$\tilde{\gamma}_i = \gamma_i \cdot M_{\nu}^{\alpha}(\delta_{i,\nu}) \cdot M_{\nu}^{\alpha}(\delta_{i+1,\nu}) \quad (14)$$

and accumulate the individual measures for an overall descriptor by taking the logarithm of the mean

$$D_{\text{ordered-chaotic}} = \log M_i^1(\tilde{\gamma}_i) \quad (15)$$

3.5 Descriptor for *smooth-coarse*

It is intuitive to identify the notion of coarseness with rough changes in loudness resp. individual frequency bands. We therefore calculate the absolute differences along the time axis and integrate along the frequency axis. For that we use a generalized mean with a low exponent ξ , thereby squeezing the magnitude differences between the individual frequency bands.

$$\delta_t = M_f^{\xi} \left| \Delta_t \hat{c}_{t,f} \right| \quad (16)$$

Then we integrate along the time axis, using a generalized mean with a tunable exponent η which allows to adjust the contrast of strong transients vs. smooth regions. We take the logarithm of the resulting value.

$$D_{\text{smooth-coarse}} = \log M_t^{\eta}(\delta_t) \quad (17)$$

3.6 Descriptor for *tonal-noisy*

The notion of pitchedness is commonly expressed by the presence of strong, isolated and stationary spectral components. This is opposed to a spectral continuum fluctuating in time, indicating noise.

To extract such qualities, we first integrate along the time axis, using a generalized mean with a relatively low exponent ξ , squeezing the magnitude range of the coefficients $\hat{c}_{t,f}$.

$$\beta_f = M_t^{\xi}(\hat{c}_{t,f}) \quad (18)$$

The we integrate along the frequency axis, using a generalized mean with a very large exponent η , considerably boosting the magnitude range of the coefficients β_f . This emphasizes strong frequency components that stick out of the noise continuum. We take the logarithm of the resulting value.

$$D_{\text{tonal-noisy}} = \log M_f^{\eta}(\beta_f) \quad (19)$$

3.7 Descriptor for *homogeneous-heterogeneous*

Here, we are interested in comparing the structural properties of the time-frequency components on a larger time-scale. In previous research [26] we have used a technique termed *fluctuation patterns* to model timbral and micro-temporal properties of textural sound. This is achieved by taking the STFT of the sonogram along the time axis.

Hence, we slice the sonogram \hat{c} into pieces s_i of uniform duration n with a *hop size* of m by use of a rectangular (boxcar) window function w_n :

$$s_{i,t,f} = \sum_{t'} w_n(t' - im) \hat{c}_{t',f} \quad (20)$$

and normalize each slice by a mean value with variable exponent ξ (probably close to 2, equaling RMS), thereby facilitating later comparison of the individual slices:

$$\bar{s}_{i,t,f} = \frac{s_{i,t,f}}{M_{t',f'}^{\xi}(s_{i,t',f'})} \quad (21)$$

We take the Discrete Fourier transform \mathcal{F} along the time axis of each slice \bar{s}_i . We are only interested in the magnitudes, representing the strengths of temporal fluctuation frequencies ρ in spectral frequency bands f .

$$\hat{s}_{i,\rho,f} = \left| \mathcal{F}_t(\bar{s}_{i,t,f}) \right| \quad (22)$$

We expect the perception of fluctuation strength to be dependent on the frequencies ρ (see [27]). Therefore, for this descriptor we skip spectral attenuation (Equation 4), but instead similarly apply interpolated attenuation factors Ψ respective to the fluctuation frequencies:

$$\tilde{s}_{i,\rho,f} = \hat{s}_{i,\rho,f} \text{ att}_{\Psi}(\rho) \quad (23)$$

By calculating the distances between consecutive instances we obtain a measure for the similarity of the slices. We use a tunable exponent η to adjust the sensitivity,

$$\delta_i = M_{\rho,f}^{\eta} |\tilde{s}_{i+1,\rho,f} - \tilde{s}_{i,\rho,f}| \quad (24)$$

and take the logarithm of the resulting value.

$$D_{\text{homogeneous-heterogeneous}} = \log M_i^1(\delta_i) \quad (25)$$

3.8 Determination of descriptor parameters

For each of the descriptors we can start out with intuitively chosen values for the variable parameters. As already mentioned above some of the parameters have been chosen with a specific numerical range in mind, e.g., low or high

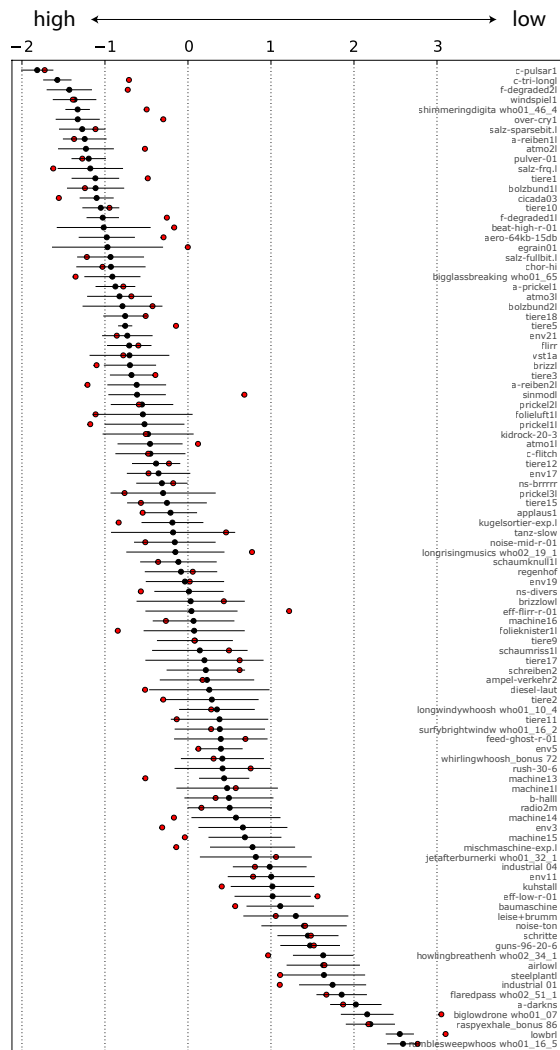


Figure 2: Comparison of values for user ratings and computed features for the construct *high–low* and 100 textural sounds. Both user ratings and computed values are whitened with respect to mean and variance. The sounds are ordered by increasing user ratings, the dots indicate mean values for each item, the horizontal bars respective standard deviations.

exponents for the generalized means. Filter coefficients can be set to zero initially. In order to arrive at optimal settings we have used a Nelder-Mead simplex optimization algorithm⁶ to tune the parameters, with the objective of high correlation with the ratings given by human listeners.

4. RESULTS

4.1 Individually optimized descriptors

Figure 2 shows user ratings (black dots) and resulting descriptor values (red dots) for the construct *high–low*. The horizontal bars centered around the black dots indicate standard deviations of the user ratings. It is clear that ratings with low standard deviations are more significant

⁶ <http://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.fmin.html>, retrieved 2012-04-11.

Construct	r weighted	r unweighted	Mean z-score
high–low	0.896	0.870	0.862
ordered–chaotic	0.742	0.677	1.259
smooth–coarse	0.751	0.725	1.005
tonal–noisy	0.750	0.696	1.019
homogeneous–heterogeneous	0.754	0.727	1.028

Table 3: Pearson correlations r (weighted and unweighted) and mean z-score for individually tuned descriptors with respect to user ratings for 100 textural sounds.

Construct	Mean z-score
high–low	0.856
ordered–chaotic	1.197
smooth–coarse	0.957
tonal–noisy	1.033
homogeneous–heterogeneous	0.944

Table 4: Mean z-score of the resulting descriptor values in relation to user ratings for a ten-fold cross validation (repeated five times) on the corpus of 100 textural sounds.

for the evaluation than those with high deviations. Therefore, for the objective function of the optimization we use a modified Pearson correlation⁷ allowing to factor in individual weights. We weight with the inverse standard deviation of each rating.

Table 2 lists the individual descriptor parameters (see Section 3) resulting from the optimization process performed on the corpus of 100 textural sounds. Some of the descriptor parameters (especially the integer ones) have been fixed to meaningful values in order to facilitate and accelerate the optimization process.

These are used to calculate the correlations achieved with respect to the human ratings, shown in Table 3, listing Pearson correlations (both weighted and unweighted) and mean z-scores. The latter expresses the error in units of the respective standard deviation.

To evaluate the robustness of the parameter fitting for sounds outside the training data, we perform a ten-fold cross-validation (parameters repetitively trained on 90% of the data, evaluated on the remaining 10%), repeated five times. Table 4 shows the results expressed in mean z-scores. The results are very similar (mostly even superior) to the z-scores achieved on the training data (see Table 3).

We are not only interested in the achievable accuracy of the individual descriptors but also in a certain amount of independence between the descriptors. Figure 3 shows a matrix of weighted Pearson cross-correlations between the individual computable descriptors and the respective qualities as rated by human listeners. As desired, the diagonal of the correlation matrix dominates the other values, indicating that qualities as perceived by listeners are relatively unambiguously modeled by the computed features. Nevertheless, there are considerable side correlations,

⁷ <http://www.mathworks.com/matlabcentral/fileexchange/20846>, retrieved 2012-04-05

Construct	Fixed parameters	Tunable parameters
high–low		$\Phi = [+8.3 \text{ dB}, +6.1 \text{ dB}], \xi = 5.24, \eta = 0.35$
ordered–chaotic	$m = 35, n = 100, \nu_{\max} = 150, \sigma = 5$	$\xi = -0.474, \eta = 100, \alpha = -100$
smooth–coarse		$\Phi = [-8.2 \text{ dB}, +16.7 \text{ dB}], \xi = 0.645, \eta = 0.91$
tonal–noisy		$\Phi = [+6.43 \text{ dB}, 0 \text{ dB}], \xi = 0.642, \eta = 100$
homogeneous–heterogeneous	$m = 35, n = 100$	$\xi = 2.48, \Psi = [+29.7 \text{ dB}, -6.1 \text{ dB}], \eta = 1.72$

Table 2: Descriptor parameters resulting from optimization with respect to high individual accuracy of the descriptors.

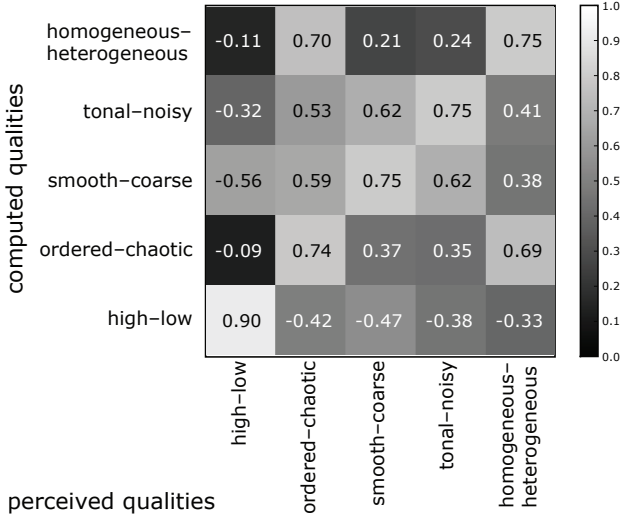


Figure 3: Weighted Pearson correlations between computed descriptors (rows), optimized for high individual accuracy, and user ratings (columns), considering 100 textural sounds. The smallest significant correlation value (at $\alpha = 0.05$, two-tailed) is ± 0.20 .

most prominently between the qualities *tonal–noisy* and *smooth–coarse*, and *ordered–chaotic* and *homogeneous–heterogeneous*. Referring to Figure 1 it can be seen that these cross-correlations are already present in the user ratings, with the same pattern, but to a smaller extent.

4.2 Optimization considering cross-correlations

Apart from the goal to optimize each descriptor separately, we can employ a global optimization scheme optimizing for the contrast between the diagonal elements and the off-diagonal elements which we wish to be as high as possible. The objective function we use strives for two goals: Firstly, to keep the main correlations as high as possible, and secondly, to increase the difference between the diagonal elements and the largest off-diagonal elements.

For each of the rows and columns of the matrix, respectively, we calculate the difference between the diagonal element and the largest absolute off-diagonal elements (by calculating M^P with large positive p , e.g. 10), plus 1 to avoid negative values. This is multiplied by the diagonal element itself. The product should be as high as possible. We will optimize for the mean (generalized) of all the inverse products with a high exponent, therefore focusing on

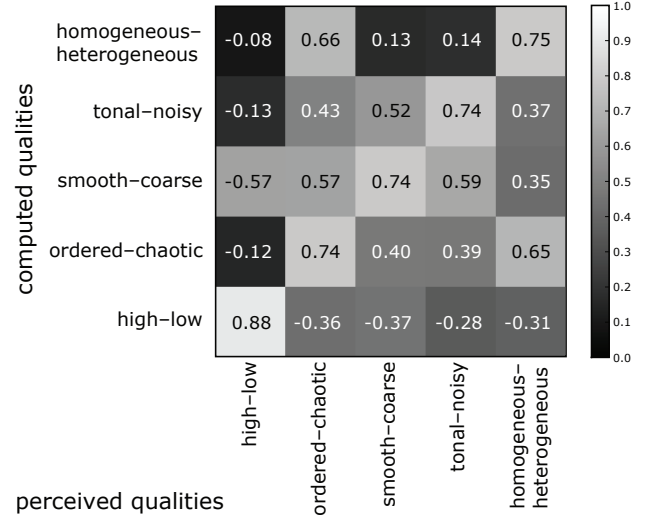


Figure 4: Weighted Pearson correlations between computed descriptors (rows), optimized for high independence, and user ratings (columns) considering 100 textural sounds. The smallest significant correlation value (at $\alpha = 0.05$, two-tailed) is ± 0.20 .

larger elements.

$$\min f(\mathbf{r}) = \prod_i \left(\left(r_{i,i} \left(r_{i,i} + 1 - \frac{M^P(|r_{i,j}|)}{j \neq i} \right) \right)^{-1} ; \left(r_{i,i} \left(r_{i,i} + 1 - \frac{M^P(|r_{j,i}|)}{j \neq i} \right) \right)^{-1} \right) \quad (26)$$

Table 4 shows the Pearson correlations between rated and computed qualities after optimization in this manner. While the main correlations (diagonal elements) decrease slightly, the differences to the larger-magnitude off-diagonal elements increase considerably. Nevertheless, the main factor for the successful modeling are the algorithms employed, and to a much lesser extent the tunable parameters. Hence, the gains achievable by such tuning are limited.

5. APPLICATION

As a proof of concept, a prototypical sound browser application⁸ has been developed which provides adequate visualization of the perceptual qualities [28] under examination, by using computed descriptor values for the corpus of 100 textural sounds. Figure 5 shows a screenshot of

⁸ <http://grrrr.org/test/texvis/map.html>, retrieved 2012-04-11. HTML5 capable web browser expected.

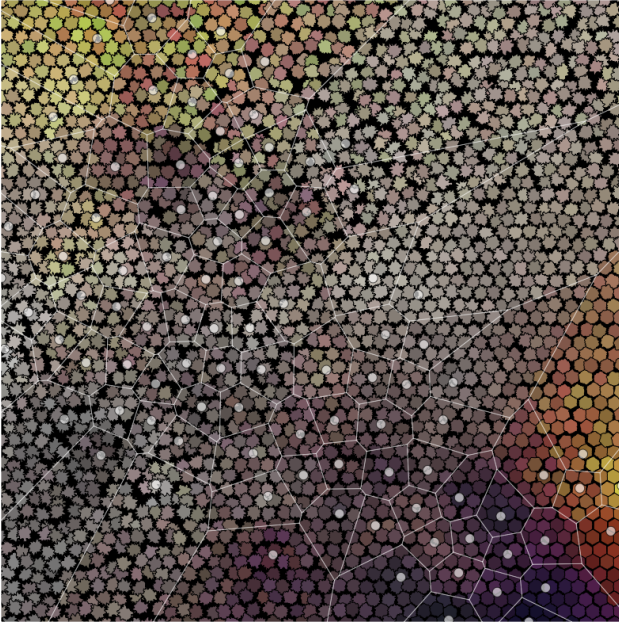


Figure 5: Interactive tiled map for browsing textural sounds, using perceptual descriptors and metaphoric, cross-modal feature visualization. Taken from [28].

this sound browser, consisting of a dynamically drawn map with white dots marking the positions of the individual sounds and a continuous tiling of the 2D space corresponding to the perceptual qualities in the map. Dimensionality reduction (from five to two dimensions) has been performed in a preprocessing step using t-SNE (t-Distributed Stochastic Neighbor Embedding, see [29]), yielding clusters of resembling perceptual qualities. These areas are perfectly reflected by the graphical representation – please note the dark and light regions, or areas with more colorful or irregularly spaced elements. Inverse Distance Weighting is used to interpolate the qualities in the areas between the sounds. A k-d-tree [30] allows for efficient retrieval of sounds to play them interactively by mouse hovering.

6. CONCLUSIONS AND OUTLOOK

We have detailed the construction of audio descriptors capable of modeling high-level, metaphoric qualities of textural sound which have been identified as perceptually relevant in previous research. Each of the descriptors contains a small number of adjustable parameters which have been tuned to a corpus of 100 textural – mostly abstract and environmental – sounds. Evaluation has yielded Pearson correlations between the audio descriptors and human ratings obtained from listening tests of above 0.74 for the constructs *ordered–chaotic*, *smooth–coarse*, *tonal–noisy*, *homogeneous–heterogeneous*, and up to 0.90 for the construct *high–low*. The descriptors are robust with respect to data external to the training corpus, as proven by ten-fold cross-validation. The resulting z-scores on test data are very similar to the ones obtained on the training data. Apart from tuning for optimal individual accuracy of the descriptors, we have also shown a strategy for the optimization with respect to enhanced independence. While

the main correlations hardly degrade, the side correlations are noticeably reduced.

So far we have only used simple algorithms and a very small number of adjustable parameters. We are convinced that with some more refinement the accuracy of our descriptors can be further increased. We have also shown the use-case for a sound browser interface using the five metaphoric audio descriptors for the intuitive visualization of sound qualities.

As of now, the calculation of the descriptors uses an off-line procedure analyzing entire sonograms. However, as this is not a fundamental necessity, we will head for a real-time capable version, based on a block-wise implementation of the NSGT algorithm [31].

In future research we will look into the possibilities of novel time-frequency representations, in particular the *scattering transform* [32], which we would like to combine with strategies of *automatic descriptor modeling* [22].

Acknowledgments

This research is supported by the Vienna Science, Research and Technology Fund (WWTF) project Audio-Miner (MA09-024). Many thanks to the anonymous reviewers for their constructive comments!

7. REFERENCES

- [1] G. Strobl, G. Eckel, and D. Rocchesso, “Sound texture modeling: A survey,” in *Proceedings of the 2006 Sound and Music Computing (SMC) International Conference*, 2006.
- [2] T. Grill, A. Flexer, and S. Cunningham, “Identification of perceptual qualities in textural sounds using the repertory grid method,” in *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, ser. AM ’11. New York, NY, USA: ACM, 2011, pp. 67–74.
- [3] P. Herrera, X. Serra, and G. Peeters, “Audio descriptors and descriptor schemes in the context of MPEG-7,” in *Proceedings of ICMC*, 1999.
- [4] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the CUIDADO project,” IRCAM, Tech. Rep., 2004.
- [5] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, July 2002.
- [6] K. Seyerlehner, G. Widmer, and T. Pohle, “Fusing block-level features for music similarity estimation,” in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010.
- [7] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, “A survey of audio-based music classification and annotation,” *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, April 2011.

- [8] J.-J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.
- [9] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK, September 2005, pp. 628–633.
- [10] T. Li and M. Ogihara, "Detecting emotion in music," in *4th International Conference on Music Information Retrieval (ISMIR 2003)*, Baltimore, Maryland, USA, October 2003.
- [11] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "State of the art report: Music emotion recognition: A state of the art review," in *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 2010, pp. 255–266.
- [12] A. Huq, J. P. Bello, and R. Rowe, "Automated music emotion recognition: A systematic evaluation," *Journal of New Music Research*, vol. 39, no. 3, pp. 227–244, May 2010.
- [13] T. Pohle, E. Pampalk, and G. Widmer, "Evaluation of frequently used audio features for classification of music into perceptual categories," in *In Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05)*, 2005.
- [14] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," in *9th International Conference on Music Information Retrieval (ISMIR 2008)*, 2008, pp. 462–467.
- [15] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, December 1980.
- [16] L. F. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition and Emotion*, vol. 12, no. 4, pp. 579–599, 1998.
- [17] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, February 2008.
- [18] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
- [19] M. Leman, V. Vermeulen, L. D. Voogdt, D. Moelants, and M. Lesaffre, "Prediction of musical affect using a combination of acoustic structural cues," *Journal of New Music Research*, vol. 34, no. 1, pp. 39–67, June 2005.
- [20] K. F. MacDorman and S. O. C. Ho, "Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison," *Journal of New Music Research*, vol. 36, no. 4, pp. 281–299, 2007.
- [21] B. Han, S. Ho, R. B. Dannenberg, and E. Hwang, "SMERS: Music emotion recognition using support vector regression," in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, October 2009, pp. 651–656.
- [22] F. Pachet and P. Roy, "Analytical features: a knowledge-based approach to audio feature generation," *EURASIP J. Audio Speech Music Process.*, vol. 2009, pp. 1:1–1:23, 2009.
- [23] P. Schaeffer, *Traité des objets musicaux*. Editions du Seuil, Paris, 1966.
- [24] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible Constant-Q transform with nonstationary gabor frames," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx 11)*, Paris, France, 2011.
- [25] A. Gutschalk, R. D. Patterson, A. Rupp, S. Upenkamp, and M. Scherg, "Sustained magnetic fields reveal separate sites for sound level and temporal regularity in human auditory cortex," *NeuroImage*, vol. 15, pp. 207–216, 2002.
- [26] T. Grill, "Re-texturing the sonic environment," in *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*, ser. AM '10. New York, NY, USA: ACM, 2010, pp. 6:1–6:7.
- [27] H. Fastl, "Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise," *Hearing Research*, vol. 8, no. 1, pp. 59 – 69, 1982.
- [28] T. Grill and A. Flexer, "Visualization of perceptual qualities in textural sounds," in *Proceedings of the International Computer Music Conference (ICMC 2012)*, Ljubljana, Slovenia, September 2012, forthcoming.
- [29] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.
- [30] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.
- [31] Holighaus, N., Dörfler, M., Velasco, G. A., and Grill, T., "The Sliced Constant-Q Transform," 2012, in preparation.
- [32] J. Andén and S. Mallat, "Multiscale scattering for audio classification," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, 2011, pp. 657–662.