



INTER-NOISE 2007

28-31 AUGUST 2007

ISTANBUL, TURKEY

Differences in the cognition of urban soundscapes and polyphonic music: a pattern recognition point of view

Jean-Julien Aucouturier^a and Boris Defreville^b

^a Ikegami Laboratory, Graduate School of Arts and Science,
The University of Tokyo, Tokyo, Japan

^b Orelia, Fontainebleau, France

ABSTRACT

This paper^c proposes to examine the differences between urban soundscapes and polyphonic music audio signals with respect to their modelling with pattern recognition algorithms (specifically, so-called "BOF" models). Such algorithms make an implicit assumption about the perceptive relevance (or saliency) of sound events, which is modelled as their statistical typicality. This is a very crude low-level cognitive model, both to model pre-attentive weighting and higher-level cognitive processes of selective attention. In this paper, we explicitly examine this model by introducing a transformation of statistical homogeneity. As expected, the BOF model of auditory saliency does not hold well for polyphonic music signals. However, surprisingly, we find that it is an efficient/sufficient representation in the case of soundscapes: frames are found to contribute to the precision of the simulated perceptive task in degrees correlated with their global statistical typicality, and overall BOF provide near-perfect replication of human judgments. These results suggest that the human perception of music and soundscapes rely on cognitive processes of a different nature: while there are important and undisputed high-level cognitive processes in soundscape perception (as supported by a wealth of recent psychological evidence), these may be less critical in shaping the overall perceptive categories than for polyphonic music. Discarding such processes hurts the (simulated) perception of music more than that of soundscapes. These findings now have to be validated by psychological experimentation.

1 INTRODUCTION

A typical approach to computer pattern recognition of audio signals represents signals as the long-term statistical distribution of local features vectors. This approach has been nicknamed "bag-of-frames" (BOF), in analogy with the "bag-of-words" (BOW) treatment of text data as a global distribution of word occurrences without preserving their organization in phrases, traditionally used in Text Classification and Retrieval. The signal is cut into short overlapping frames (typically 50ms with a 50% overlap), and for each frame, a feature vector is computed. Features usually consist of a generic, all-purpose spectral representation such as

^a Email address: aucouturier@gmail.com

^b Email address: defreville@orelia.fr

^c Parts of the results reported here have been published in Aucouturier, J.-J., Defreville, B. and Pachet, F. The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. Journal of the Acoustical Society of America, 2007.

Mel Frequency Cepstrum Coefficients (MFCC) [1]. The physical source of individual sound samples is not explicitly modelled: all feature vectors are fed to a classifier (based e.g. on Gaussian Mixture Models [2]) which models the global distributions of the features of signals corresponding to each class. Global distributions for each class can then be used to compute decision boundaries between classes. A new, unobserved signal is classified by computing its feature vectors, finding the most probable class for each of them, and taking the overall most represented class for the whole signal.

The BOF approach has proved very effective for soundscapes. Ma et al. [3] report 91% classification precision on a database of 80 3-second sound extracts from 10 everyday soundscape classes (street, factory, football game, etc.).

For the analysis of polyphonic music signals also, the BOF approach has led to some success and is by far the most predominant paradigm. Each contribution typically instantiates the same basic architecture described above, only with different algorithm variants and parameters. Although they use the same underlying rationale of modelling global timbre/sound in order to extract high-level descriptions, the spectrum of the targeted descriptions is rather large: genre, mood, singing language to name but a few (see [4] for a review).

However, contrary to its application to soundscapes, recent research on the issue of polyphonic timbre similarity shows that BOF seems to be bounded to moderate performance, most notably:

- Glass ceiling: Surprisingly, thorough exploration of the space of typical algorithms and variants (such as different signal features, static or dynamic models, parametric or non-parametric estimation, etc.) and exhaustive fine-tuning of the corresponding parameters fail to improve the precision above a empirical glass-ceiling, around 70% precision (although this of course should be defined precisely and depends on tasks, databases, etc.) [4].
- Paradox of dynamics: Further, traditional means to model data dynamics, such as delta-coefficients, texture windows or Markov modelling, do not provide any improvement over the best static models for real-world, complex polyphonic textures of several seconds length [5]. This is a paradoxical observation, since static models consider all frame permutations of the same audio signal as identical, while this has a critical influence on their perception. Moreover, psychophysical experiments have established the importance of dynamics, notably the attack time and fluctuations of the spectral envelope, in the perception of individual instrument notes.
- Hubs: Finally, recent experiments [6] show that the BOF approach (when used on polyphonic music) tends to create false positives which are mostly always the same songs regardless of the query. In other words, there exist songs, which we have called hubs, which are irrelevantly close to all other songs.

One of the possible explanations for the limitations of the BOF approach is that it makes a very crude model for the perception of an auditory signal. Distributions are compared (e.g. with the Kullback Leibler distance) on the basis of their most stereotypical frames. Therefore, with BOF algorithms, frames contribute to the simulation of the auditory sensation in proportion of their statistical predominance in the global frame distribution.

In other words, the *perceptive saliency* of sound events is modelled as their *statistical typicality*.

BOF is not intended (neither here nor in the pattern recognition literature) as a cognitive model, but rather is an engineering technique to simulate and replicate the outcome of the corresponding human processing. Nevertheless, it is useful to note that the above model of auditory saliency would be a very crude cognitive model indeed, both to model

- pre-attentive weighting (which has been found a correlate of frequency and temporal contrasts, i.e. arguably the exact opposite of statistical typicality) [7]
- and higher-level cognitive processes of selective attention (which are partly under voluntary control, hence products of many factors such as context and culture) [8]

In this paper, we propose to examine explicitly this saliency hypothesis by introducing a specially-designed “statistical homogeneity” transform to audio signals, which only keeps frames in the signal which are the most statistically prototypical of the overall distribution. We study the influence of each transform on the precision of BOF modelling for 2 datasets, one composed of polyphonic music pieces, and the other of sound recordings of urban environments (soundscapes). We observe very different behaviors. It appears that, contrary to environmental textures, not all music frames are equally discriminative: minority frames (the 5% less statistically significant ones) are extremely important for music while they can be discarded to notable advantage for soundscapes. Moreover, it appears that there exists, in typical polyphonic music distributions, a population of frames (in the range [60%-90%] of statistical weight) which is detrimental to the modelling of perceptual similarity.

2 ALGORITHM AND DATASETS

We study the influence of a statistical homogeneity transform applied to a pattern recognition algorithm, which is an implementation of the BOF approach. We describe here the BOF algorithm (in fact a similarity measure) as well as the two datasets used in the experiments. Then we introduce a homogeneity transform that can be applied on the canonical similarity measure to derive new similarity measures with varying homogeneity. We finally describe how to measure the precision of such homogenized similarity measures. Results of the experiments are reported in Section 3.

2.1 Algorithm

We sum up here the audio similarity algorithm presented in [4]. The signal is first cut into frames. For each frame, we estimate the spectral envelope by computing a set of Mel Frequency Cepstrum Coefficients (MFCCs). The Cepstrum is the inverse Fourier transform of the logarithm of the Fourier spectrum $\log S$.

$$c_n = \frac{1}{2\pi} \int_{\omega=-\pi}^{\omega=\pi} \log S(\omega) \exp j\omega n \, d\omega$$

We call Mel-Cepstrum the Cepstrum computed after a non-linear frequency warping onto a perceptual frequency scale, the Mel-frequency scale [10], which reproduces the non-linearity of the frequency resolution of the human auditory system (low Hertz frequencies are more easily discriminated than high Hertz frequencies). The c_n in equation are called Mel frequency cepstrum coefficients (MFCCs), of which we keep a given number N .

We then model the distribution of the MFCCs over all frames using a Gaussian Mixture Model (GMM). A GMM estimates a probability density as the weighted sum of M simpler Gaussian densities, called components or states of the mixture:

$$p(x_t) = \sum_{m=1}^{m=\mathcal{M}} \pi_m \mathcal{N}(x_t, \mu_m, \Sigma_m)$$

where x_t is the feature vector observed at time t , \mathcal{N} is a Gaussian pdf with mean μ_m , covariance matrix Σ_m , and π_m is a mixture coefficient (also called state prior probability). The parameters of the GMM are learned with the classic E-M algorithm [2].

We then compare the GMM models to match different signals, which gives a similarity measure based on the audio content of the items being compared. We use a Monte Carlo approximation of the Kullback-Leibler (KL) distance between each duple of models A and B. The KL-distance between 2 GMM probability distributions p_A and p_B (as defined above) is defined by:

$$d(A, B) = \int p_A(x) \log \frac{p_B(x)}{p_A(x)} dx$$

The KL distance can thus be approximated by the empirical mean :

$$\widetilde{d(A, B)} = \frac{1}{n} \sum_{i=1}^n \log \frac{p_B(x_i)}{p_A(x_i)}$$

(where n is the number of samples x_i drawn according to p_A) by virtue of the central limit theorem.

In this work, we use the optimal settings determined by previous research in the context of polyphonic music [4], namely 20 MFCCs appended with 0th order coefficient, 50-component GMMs, compared with $n = 2000$ Monte-Carlo draws.

2.2 Datasets

2.2.1. Urban soundscapes

For this study, we gathered a database of 106 3-minute recordings of urban soundscapes, recorded in Paris using a omni-directional microphone. The recordings are clustered in 4 “general classes”:

- Avenue: Recordings made on relatively busy thoroughfares, with predominant traffic noise, notably buses and car horns.
- Neighborhood: Recordings made on calmer neighborhood streets, with more diffuse traffic, notably motorcycles, and pedestrian sounds.
- Street Market: Recordings made on street markets in activity, with distant traffic noise and predominant pedestrian sounds, conversation and auction shouts.
- Park: Recordings made in urban parks, with lower overall energy level, distant and diffuse traffic noises, and predominant nature sounds, such as water or bird songs.

Recordings are further labeled into 11 “detailed classes”, which correspond to the place and date of recording of a given environment. For instance, “Parc Montsouris (Paris 14e)” is a subclass of the general “Park” class. Some detailed classes also discriminate takes at identical locations and dates, but with some exceptional salient difference. For instance, “Marche Richard Lenoir (Paris 11e)” is a recordings made in a street market on Boulevard Richard Lenoir in Paris, and “Marche Richard Lenoir (music)” is a recording made on the same day of the same environment, only with the additional sound of a music band playing in the street. Table I shows the details of the classes used, and the number of recordings available in each class.

Table 1: Composition of the urban soundscape dataset

Class	Detailed Class	Size
Avenue	Boulevard Arago	14
Avenue	Boulevard du Trone	5
Avenue	Boulevard des Marchaux	8
Street	Rue de la Sant	7
Street	Rue Reille day1	14
Street	Rue Reille day2	7
Market	Marché Glacière	8
Market	Marché R. Lenoir	22
Market	Marché R. Lenoir (music)	9
Park	Parc Montsouris Spring	20
Park	Parc Montsouris Summer	8

2.2.2 Polyphonic Music

The polyphonic music dataset used in this study contains 350 popular music titles. It is organized in 37 clusters of songs by the same artist, encompassing very different genres and instrumentations (from Beethoven piano sonata to The Clash’s punk-rock or Musette-style accordion). Artists and songs were chosen in order to have clusters that are “timbrally” consistent (all songs in each cluster sound the same). Furthermore, they only include songs that are timbrally homogeneous, i.e. there is no big texture change within each song. The database is constructed so that nearest neighbors of a given song should optimally belong to the same cluster as the seed song. Details on the design and contents of this database can be found in [4].

2.3 Statistical homogenization transform

We define a statistical homogeneity transform $hk: G \rightarrow G$ on the space G of all GMMs, where k in $[0, 1]$ is a percentage value, as:

```

 $g_2 = h_k(g_1)$ 
 $(c_1, \dots, c_n) \leftarrow \text{sort}(\text{components}(g_1), \text{decreasing } w_c)$ 
define  $\mathcal{S}(i) = \sum_{j=1}^i \text{weight}(c_j)$ 
 $i_k \leftarrow \arg \min_{i \in [1, n]} \{\mathcal{S}(i) \geq k\}$ 
 $g_2 \leftarrow \text{newGMM}(i_k)$ 
define  $d_i = \text{component}(g_2, i)$ 
 $d_i \leftarrow c_i, \forall i \in [1, i_k]$ 
 $\text{weight}(d_i) \leftarrow \text{weight}(c_i) / \mathcal{S}(i_k), \forall i \in [1, i_k]$ 
return  $g_2$ 
end  $h_k$ 

```

From a GMM g trained on the total amount of frames of a given song, the transform h_k derives an homogenized version of g which only contains its top $k\%$ components. Frames are all the more so likely to be generated by a given gaussian component c than the weight w_c of the component is high (w_c is also called prior probability of the component). Therefore, the homogenized GMM accounts for only a subset of the original song's frames: those that amount to the $k\%$ most important statistical weight. For instance, $h_{99\%}(g)$ creates a GMM which doesn't account for the 1% least representative frames in the original song.

We apply 11 transforms h_k for k in $[20, 40, 60, 80, 90, 92, 94, 96, 98, 99, 100]$ to the GMMs used in the similarity measure described above. Each transform is applied on each dataset, thus yielding two sets of 11 similarity measures, which we can compare using the metric below.

2.4 Evaluation metric

Similarity measures derived by applying the above transform to the canonical algorithm described in 2.1 are compared by computing their precision after 5, 10 and 15 documents are retrieved, and their R-precision, i.e. their precision after all relevant document are retrieved. Each value measures the ratio of the number of relevant documents to the number of retrieved documents. The set of relevant documents for a given sound sample is the set of all samples of the same category than the seed. This is identical to the methodology used e.g. in [4].

3 RESULTS

3.1 Influence on variance

Figure 1 shows the influence of the statistical homogenization transform on the variance of the resulting GMM for both datasets. The variance of a GMM model can be defined by sampling a large number of points from this model, measuring the variance of these points in each dimension, and summing the deviations together. This is equivalent to measuring the norm of the covariance matrix of a single-component GMM fitted to the distribution of points. The transformation has a very distinct influence on each type of audio signals. Removing the least important 1% frames from urban soundscape signals drastically reduces the GMM variance by more than 50%. However, further statistical homogenization has little influence on the overall variance. This indicates that soundscape signals are very homogeneous and redundant statistically, except for a very small proportion of outlier frames (the least significant 1%), which account for a half of the overall variance, and probably

represent very different MFCC frames than the ones composing the main mass of the distribution. Such frames would typically represent very improbable sound events which are not characteristic of a given environment, such as the occasional plane flying over a park.

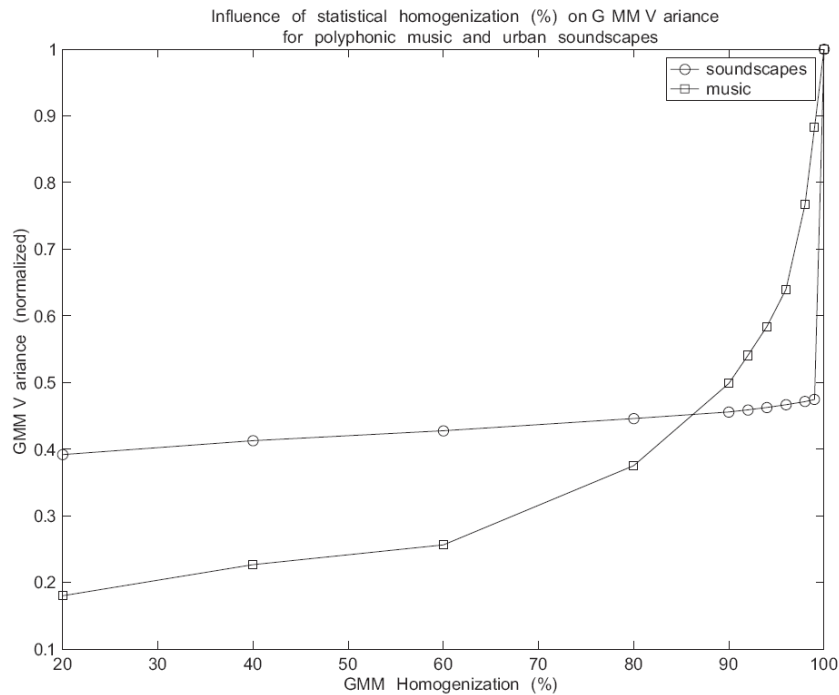


Figure 1: Influence of statistical homogeneity transform on the variance of the GMMs of urban soundscapes and music signals

When applied to polyphonic music signals, it appears that the homogenization transform reduces the variance of the models exponentially. Half of the original variance is explained by the 10% least representative frames and more than 80% by the 40% least representative frames. This indicates a greater heterogeneity than for soundscape signals, and a more diffuse notion of “outlier” frames.

3.2 Influence on precision

Figure 2 shows the influence of statistical homogenization on the precision of the resulting similarity measure, for both datasets. The precision for urban soundscapes is measured with the 10-precision using the detailed classes as ground truth, and with the R-precision for polyphonic music. For both dataset the precision is measured by reference to the baseline precision corresponding to $k = 100\%$, which is greater for soundscapes than music as already discussed.

On both datasets, increased homogenization decreases the precision of the similarity measure: homogenization with $k = 20\%$ degrades the measure’s precision by 6% (relative) for urban soundscapes, and by 17% (relative) for polyphonic music. It seems reasonable to interpret the decrease in precision when k decreases as a consequence of reducing the amount of discriminative information in the GMMs (e.g. from representing a given song, down to a more global style of music, down to the even simpler fact that it is music). Apart from this general trend however, the transform has a very different influence on the measure’s precision depending on the class of audio signals.

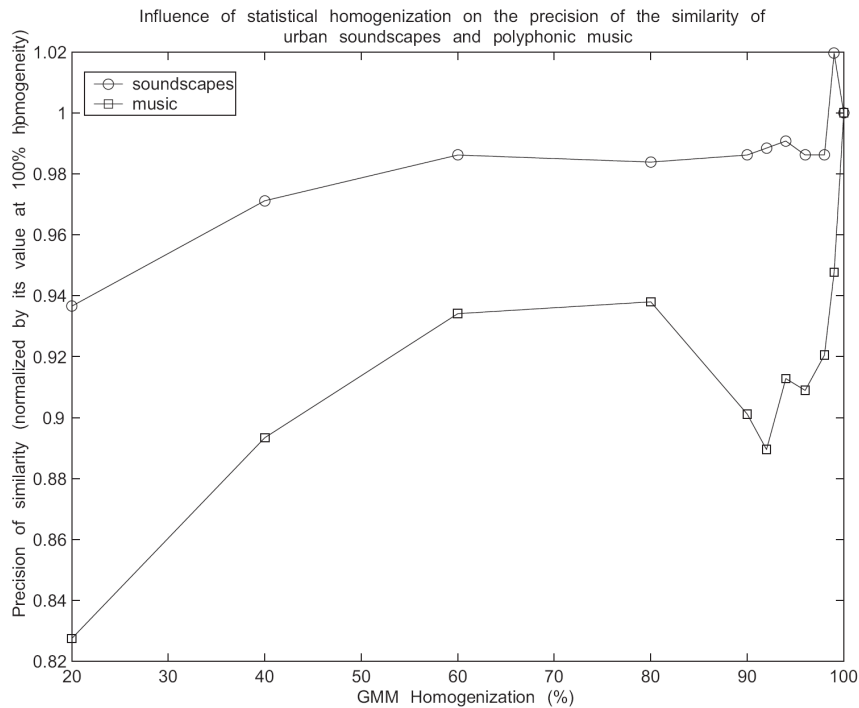


Figure 2: Comparison of the influence of statistical homogeneity transform on the precision of the similarity measure for urban soundscapes and music signals.

In the case of urban soundscapes, 99% homogenization is slightly beneficial to the precision. This suggests that the 1% less significant frames, which were found in Figure 1 to account for half of the overall variance, are spurious frames which are worth smoothing out. Further homogenization down to 60% has a moderate impact on the precision, which is reduced by about 1% (absolute). The decrease in precision from 99% down is monotonic. This suggests that the frame distribution from 99% down is very homogeneous and redundant. Urban soundscapes can be discriminated nearly optimally by considering only the most significant 50% of the frames.

In the case of polyphonic music, the decrease in precision is not monotonic. Figure 2 clearly shows a very important decrease in the precision in the first few percent of homogenization. The severely degraded precision observed for $k = 30\%$ is reached as early as $k = 95\%$. This is a strong observation: the precision of the measure seems to be controlled by an extremely small amount of critical frames, which represent typically less than 5% of whole distribution. Moreover, these frames are the least statistically significant ones, i.e. are modelled by the least important gaussian components in the GMMs. This indicates that the majority (more than 90%) of the MFCC frames of a given song are a very poor representation of what discriminates this song from other songs. This is the exact opposite behavior to the one observed for soundscape signals, where these least significant frames can be removed to some advantage.

Moreover, Figure 2 shows that after the abrupt sink when removing the first 5% frames in typical music distributions, the precision tends to increase when k decreases from 90% to 60%, and then decreases again for k smaller than 60%. The maximum value reached between 60% and 80% is only 6% (relative) lower than the original value at $k = 100\%$. The behavior in Figure 2 suggests that there is a population of frames in the range [60%, 95%] which is

mainly responsible for the bad precision of the measure on music signals. While the precision of the measure increases as more frames are included when k increases from 20% to 60% (such frames are increasingly specific to the song being modelled), it suddenly decreases when k gets higher than 60%, i.e. this new 30% information is detrimental for the modelling and tend to diminish the discrimination between songs. The continuous degradation from 60% to 95% is only eventually compensated by the inclusion of the final 5% critical frames.

4 CONCLUSION

The above results establish, as expected, that the mechanism of auditory saliency implicitly assumed by the BOF approach does not hold for polyphonic music signals: for instance, frames in statistical minority have a crucial importance in simulating perceptive judgments. However, surprisingly, the crude saliency hypothesis seems to be an efficient/sufficient representation in the case of soundscapes: frames are found to contribute to the precision of the simulated perceptive task in degrees correlated with their global statistical typicality, and overall BOF provide near-perfect replication of human judgments.

The fact that such a simple model is sufficient to simulate the perception of soundscapes could suggest that the cognitive processes involved in their human processing are less “demanding” than for polyphonic music. This finding is only based on algorithmic considerations, and naturally would have to be validated with proper psychological experimentations. Nevertheless, it seems at odds with a wealth of recent psychological evidence stressing that soundscapes judgments doesn’t result of a low-level immediate perception, but rather high-level cognitive reasoning which accounts for the evidence found in the signal, but also depends on cultural expectations, a-priori knowledge or context. For instance, the subjective evaluation of urban soundscapes has been found to depend as much on semantic features than perceptual ones: soundscapes reflecting activities with higher cultural values (e.g. human vs mechanical) are systematically perceived as more pleasant [9]. Similarly, cognitive categories have been found to be mediated by associated behaviors and interaction with the environment: a given soundscape can be described as e.g. “too loud to talk”, but “quiet enough to sleep” [10].

What our results could indicate is that, while there are indeed important and undisputed high-level cognitive processes in soundscape perception, these may be less critical in shaping the overall perceptive categories than for polyphonic music. Discarding such processes hurts the perception of music more than that of soundscapes. A possible reason for this is that there are important specificities in the structure of polyphonic music, namely very definite temporal units (e.g. notes) with both internal (transient, steady-state) and external (phrase, rhythm) organization. Such structural specificities in polyphonic music signals may require cognitive processes active on a more symbolic and analytical level than what can be accounted for by the BOF approach, which essentially builds an amorphous and holistic description of the object being modelled. These computational experiments open the way for more careful psychological investigations of the perceptive paradoxes proper to polyphonic music timbre, in which listeners “hear” things that are not statistically significant in the actual signal, and that the low-level models of timbre similarity studied in this work are intrinsically incapable of capturing [11].

5 REFERENCES

- [1] L. Rabiner and B. Juang, *Fundamentals of speech recognition* (Prentice-Hall) (1993).
- [2] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford Press) (1995)
- [3] L. Ma, D. Smith, and B. Milner, “Context awareness using environmental noise classification”, in *Proceedings of 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland (2003).
- [4] J.-J. Aucouturier and F. Pachet, “Improving timbre similarity: How high’s the sky?” *Journal of Negative Results in Speech and Audio Sciences* 1 (2004).
- [5] J.-J. Aucouturier and F. Pachet, “The influence of polyphony on the dynamical modelling of musical timbre”, *Pattern Recognition Letters* (in press) (2007).
- [6] J.-J. Aucouturier and F. Pachet, “A scale-free distribution of false positives for a large class of audio similarity measures.”, *Pattern Recognition* (in press) (2007).
- [7] C. Kayser, C. Petkov, M. Lippert, and N. K. Logothetis, “Mechanisms for allocating auditory attention: an auditory saliency map”, *Current Biology* 15, 1943–1947 (2005).
- [8] P. Janata, B. Tillmann, and J. Bharucha, “Listening to polyphonic music recruits domain-general attention and working memory circuits”, *Cognitive, Affective and Behavioral Neuroscience* 2(2), 121–140 (2002).
- [9] D. Dubois, C. Guastavino, and M. Raimbault, “A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories”, *Acta Acustica united with Acustica* 92, 865–874 (2006).
- [10] C. Guastavino, “Categorization of environmental sounds”, *Canadian Journal of Experimental Psychology* (2007), (in press).
- [11] J.-J. Aucouturier, F. Pachet, “How much audition involved in everyday categorization of music?” (submitted) *Cognitive Science*, 2007.