The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music

Jean-Julien Aucouturier[a)]
Ikegami Lab, Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Boris Defreville[b)]
ORELIA, 77300 Fontainebleau, France

François Pachet[c)]
SONY CSL, 6 rue Amyot 75005 Paris, France

(Received 13 April 2006; revised 30 January 2007; accepted 29 May 2007)

The “bag-of-frames” approach (BOF) to audio pattern recognition represents signals as the long-term statistical distribution of their local spectral features. This approach has proved nearly optimal for simulating the auditory perception of natural and human environments (or soundscapes), and is also the most predominent paradigm to extract high-level descriptions from music signals. However, recent studies show that, contrary to its application to soundscape signals, BOF only provides limited performance when applied to polyphonic music signals. This paper proposes to explicitly examine the difference between urban soundscapes and polyphonic music with respect to their modeling with the BOF approach. First, the application of the same measure of acoustic similarity on both soundscape and music data sets confirms that the BOF approach can model soundscapes to near-perfect precision, and exhibits none of the limitations observed in the music data set. Second, the modification of this measure by two custom homogeneity transforms reveals critical differences in the temporal and statistical structure of the typical frame distribution of each type of signal. Such differences may explain the uneven performance of BOF algorithms on soundscapes and music signals, and suggest that their human perception rely on cognitive processes of a different nature. © *2007 Acoustical Society of America.* [DOI: 10.1121/1.2750160]

PACS number(s): 43.60.Cg, 43.50.Rq, 43.60.Lq, 43.66.Jh, 43.66.Ba [BSF] Pages: 881–891

I. INTRODUCTION

A. Soundscapes

In 1977, composer R. Murray Schafer coined the term *soundscape* as an auditory equivalence to landscape.[1] He proposed to consider soundscapes as musical compositions, in which the sound sources are musical instruments. Nowadays, the concept of soundscape is used as a methodological and theoretical framework in the field of rural or urban sound quality, notably for the assessment of noise annoyance.[2] Psycho-physic experiments on the perception of soundscapes[3–5] indicate that the cognitive processes of recognition and similarity operate on the basis of the identification of the physical sources. For instance, a given soundscape can be classified as a “park,” when specific and localized audio events such as “birds singing,” or “children playing” are identified.[6] This also holds for semantic categorization,[7] i.e., the subjective “unpleasantness” of urban soundscapes increases when more mechanical sound sources (e.g., vehicles) are identified than natural sources (e.g., voices or birds). However, recent research[8] shows that people are also capable of more holistic strategies for processing soundscapes, when individual source identification is difficult in the presence of too many noncharacteristic events (“background noise”).

There have been various attempts to simulate human perception of soundscapes with computer algorithms, with methodologies that closely resemble the two alternative cognitive strategies mentioned earlier. A majority of contributions[9–14] take the strategy to identify the constituent sound sources individually. The typical implementation describes sound extracts with generic frame-level features, such as MPEG-7 spectral descriptors,[11] and use hidden Markov models[15] to represent their statistical dynamics. Recent research[14] proposes to enhance this typical scheme by learning problem-specific features, adapted to each sound class, with genetic programming.

However, another trend of works[16–18] proposes to directly recognize soundscapes as a whole, without the prior identification of constituent sound sources. In these works, sound-scapes are modeled as the long-term accumulative distribution of frame-based spectral features. This approach has been nicknamed “bag-of-frames” (BOF), in analogy with the “bag-of-words” treatment of text data as a global distribution of word occurrences without preserving their organization in phrases, traditionally used in text classification and

[a)] Electronic mail: aucouturier@gmail.com
[b)] Electronic mail: boris.defreville@orelia.fr
[c)] Electronic mail: pachet@csl.sony.fr

TABLE I. Number of contributions using the bag-of-frames paradigm in past ISMIR symposiums.

Year	BOF papers	Total papers	Percentage
2000	6	26	23
2001	9	36	25
2002	14	58	24
2003	12	50	24
2004	23	104	22
2005	24	114	21
Total	88	388	23

retrieval.[19] The signal is cut into short overlapping frames (typically 50 ms 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 mel frequency cepstrum coefficients[15] (MFCC). The physical source of individual sound samples is not explicitly modeled: All feature vectors are fed to a classifier (based, e.g., on Gaussian mixture models[20]) which models the global distributions of the features of signals corresponding to each class (e.g., pedestrian street or park). 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.*[18] report 91% classification precision on a database of 80 3 s sound extracts from 10 everyday soundscape classes (street, factory, football game, etc.). Notably, such systems seem to perform better than average human accuracy on the same task (35%), which suggests that 3 s audio data provide enough information for pattern recognition, but not for people. Similarly, Peltonen *et al.*[4] report that the average recognition time for human subjects on a list of 34 soundscapes is 20 s. This supports the cognitive strategy of source identification, which typically imposes longer latencies, depending on the temporal density of discriminative sound events.

B. Music

For the analysis of polyphonic music signals also, the BOF approach has led to some success and is by far the most predominant paradigm. Table I shows an enumeration of paper and poster contributions in the ISMIR conference[21] since its creation in 2000. Each year, about a fourth of all papers, and on the whole 88 papers out of a total 388, use the approach. Each contribution typically instantiates the same basic architecture described earlier, only with different algorithm variants and parameters. Although they use the same underlying rationale of modeling global timbre/sound in order to extract high-level descriptions, the spectrum of the targeted descriptions is rather large: genre,[22] mood,[23] singing language[24] to name but a few.

However, contrary to its application to soundscapes, recent research[25–27] on the issue of polyphonic timbre similarity shows that BOF seems to be bounded to moderate performance, most notably:

(1) Glass ceiling: Surprisingly, thorough exploration of the space of typical algorithms and variants (such as different signal features, static or dynamic models, parametric or nonparametric estimation, etc.) and exhaustive fine-tuning of the corresponding parameters fail to improve the precision above an empirical *glass ceiling*,[25] around 70% precision (although this of course should be defined precisely and depends on tasks, databases, etc.).

(2) Paradox of dynamics: Further, traditional means to model data dynamics, such as delta coefficients, texture windows, or Markov modeling, do not provide any improvement over the best static models for real-world, complex polyphonic textures of several seconds length.[26] 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[28] have established the importance of dynamics, notably the attack time and fluctuations of the spectral envelope, in the perception of individual instrument notes.

(3) Hubs: Finally, recent experiments[27] 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. This phenomenon is reminiscent of other results in different domains, such as speaker recognition[29] or fingerprint identification,[30] which intriguingly also typically rely on the same BOF approach. This suggests that this could be an important phenomenon which generalizes over the specific problem of polyphonic music similarity, and indicates a general structural property of the class of algorithms examined here, at least *for a given class of signals* to be defined.

C. Objectives

This paper proposes to re-evaluate this situation and to explicitly examine the difference between soundscape and polyphonic music signals with respect to their modeling with the BOF approach.

We apply to a data set of urban soundscapes an algorithmic measure of acoustic similarity that we introduced[25] in the context of polyphonic music. The measure is a typical instantiation of the BOF approach, namely comparing the long-term distributions of MFCC vectors, using Kullback-Leibler divergence between Gaussian mixture models. For music, the measure approximates the perception of similar global timbre, e.g., of songs that "sound the same." As already noted, the measure only achieves moderate precision on music and shows notable discrepancies with human perception. We find here that the same measure is nearly optimal for modeling the perceptual similarity of urban soundscapes. This confirms the situation found in the literature that soundscape and polyphonic music signals are not equal with respect to their modeling with the BOF approach. Notably, the application of timbre similarity to soundscapes does not seem to create hubs.

To explain these differences, we report on two experiments in which we apply specially designed *homogeneity* transforms to each data sets:

(1) Temporal homogeneity, which folds an original signal onto itself a number of times, so the resulting signal only contains a fraction of the original data.
(2) Statistical homogeneity, 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 modeling for both sound-scapes and music, and show very different behaviors. This notably establishes that the distribution of frame-based spectral features is very homogeneous for soundscapes, which makes their BOF modeling very robust to data transformations. i.e., soundscapes can be compressed to only a small fraction of their duration without much loss in terms of distribution modeling. Polyphonic music on the contrary seems to require a large quantity of feature information in order to be properly modeled and compared. Furthermore, 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 modeling of perceptual similarity.

II. ACOUSTIC SIMILARITY OF URBAN SOUNDSCAPES AND POLYPHONIC MUSIC

A. Algorithm

We sum up here the timbre similarity algorithm presented in Aucouturier and Pachet (2004).[25] The signal is first cut into frames. For each frame, we estimate the spectral envelope by computing a set of MFCCs. We then model the distribution of the MFCCs over all frames using a Gaussian mixture model (GMM). GMM estimates a probability density as the weighted sum of $\mathcal{M}$ simpler Gaussian densities, called components or states of the mixture:

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

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.[20]

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 two GMM probability distributions p_A and p_B [as defined in Eq. (1)] is defined by

$$d(A,B) = \int p_A(x) \log \frac{p_B(x)}{p_A(x)} dx. \tag{2}$$

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

$$\widetilde{d(A,B)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_B(x_i)}{p_A(x_i)} \tag{3}$$

(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,[25] namely 20 MFCCs appended with zeroth-order coefficient, 50-component GMMs, compared with $n=2000$ Monte Carlo draws.

B. Data sets

1. Urban soundscapes

For this study, we gathered a database of 106 3 min recordings of urban soundscapes, recorded in Paris using an omnidirectional microphone. The recordings are clustered in four "general classes" as follows:

(1) Avenue: Recordings made on relatively busy thoroughfares, with predominant traffic noise, notably buses and car horns.
(2) Neighborhood: Recordings made on calmer neighborhood streets, with more diffuse traffic, notably motorcycles, and pedestrian sounds.
(3) Street market: Recordings made on street markets in activity, with distant traffic noise and predominant pedestrian sounds, conversation, and auction shouts.
(4) 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 14è)" is a subclass of the general "Park" class. Some detailed classes also discriminate at identical locations and dates, but with some exceptional salient difference. For instance, "Marché Richard Lenoir (Paris 11è)" is a recording made in a street market on Boulevard Richard Lenoir in Paris, and "Marché 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 II shows the details of the classes used, and the number of recordings available in each class.

2. Polyphonic music

The polyphonic music data set used in this study contains 350 popular music titles, extracted from the Cuidado database.[31] 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* punk rock and *Musette*-style accordion). Artists and songs were chosen in order to have clusters that are "timbrally" consistent (all

TABLE II. Composition of the urban soundscape database.

Class	Detailed class	Size
Avenue	Boulevard Arago	14
Avenue	Boulevard du Trône	5
Avenue	Boulevard des Maréchaux	8
Street	Rue de la Santé	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

songs in each cluster sound the same). Furthermore, we only select songs that are timbrally homogeneous, i.e., there is no big texture change within each song. The test 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 Aucouturier and Pachet (2004).[25]

C. Evaluation metric

The algorithms 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 as the seed. This is identical to the methodology used, e.g., in Aucouturier and Pachet (2004).[25]

D. Results

1. Precision

Table III gives the precision of timbre similarity applied to both data sets. It appears that the results are substantially better for urban soundscapes than for polyphonic music signals, nearing perfect precision in the first five nearest neighbors even for detailed classes. High precision using the general classes shows that the algorithm is able to match recordings of different locations on the basis of their sound level (avenues, streets), and sound quality (pedestrian, birds). High precision on detailed classes shows that the algorithm is also able to distinguish recordings of the same environment made at different times (spring or summer), or in different contexts (with and without music band). This result has a natural application to computer-based classification, e.g., using a simple k-nearest neighbor strategy, and could prove useful for context-recognition, for instance in the context of wearable computing.[32]

TABLE III. Comparison of similarity measure for urban soundscapes and polyphonic music.

Database		5 Prec.	10 Prec.	15 Prec.	R Prec.
Music		0.73	0.70	0.65	0.65
Soundscapes	General	0.94	0.87	0.77	0.66
	Detailed	0.90	0.79	0.75	0.74

TABLE IV. Five most frequent false positives in the music database.

Song	N_{10}
Mitchell, Joni - Don Juan's Reckless Daughter	57
Moore, Gary - Separate Ways	35
Rasta Bigoud - Tchatche est bonne	30
Public Enemy - Cold Lampin With Flavor	27
Gilberto, Joao - Tin tin por tin tin	25

2. Hubs

As mentioned earlier, an intriguing property of the application of the similarity measure to polyphonic music signals is that it tends to create false positives which are mostly always the same songs regardless of the query. In other words, there exist songs, which we call *hubs*, which are irrelevantly close to all other songs. We give a detailed description of this phenomenon in Aucouturier and Pachet (2007).[27]

A natural measure of the hubness of a given song is the number of times the song occurs in the first n nearest neighbors of all the other songs in the database. An important property of the number of n occurrences N_n of a song is that the sum of the values for all songs is constant given a database. Each query only gives the opportunity for n occurrences to the set of all the other songs, such that the total number of n occurrences in a given $\mathcal{N}$-size database is $n*\mathcal{N}$. Therefore, the mean n occurrence of a song is equal to n, independent of the database and the distance measure.

Table IV shows the five biggest hubs in the polyphonic music database ranked by the number of times they occur in the first ten nearest neighbors over all queries (N_{10}). This illustrates the predominance of a few songs that occur very frequently. For instance, the first song, *Mitchell, Joni—Don Juan's Reckless Daughter* is very close to 1 song out of 6 in the database (57 out of 350), which is more than six times more than the theoretical mean value (10). Among these occurrences, many are likely to be false positives.

Figure 1 shows the histogram of the number of 20-occurrences N_{20} obtained with the above-mentioned distance on the database of urban soundscapes, compared with the same measure on the test database of polyphonic music. It appears that the distribution of number of occurrences for soundscapes is more narrow around the mean value of 20, and has a smaller tail than the distribution for polyphonic music. Notably, there are four times as many audio items with more than 40 20-occurrences in the music data set than in the urban soundscape data set. This is also confirmed by the manual examination of the similarity results for the urban soundscapes: none of the (few) false positives reoccur significantly more than random.

This establishes the fact that hubs are not an intrinsic property of the class of algorithm used here, but rather appear only for a certain classes of signals, which includes polyphonic music, but not urban soundscapes.

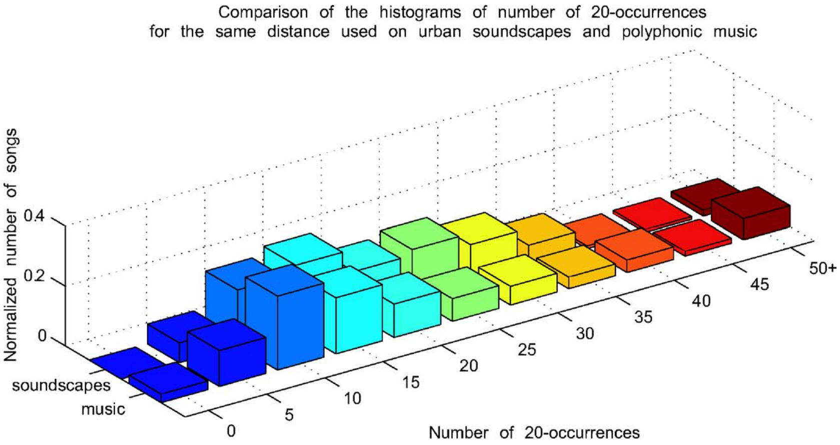

FIG. 1. (Color online) Comparison of the histograms of number of 20-occurrences for the same distance used on urban soundscapes and polyphonic music.

On the whole, these results confirm that urban soundscapes and polyphonic music signals are not equal with respect to their modeling with the BOF approach. To explain these differences, we now report on two experiments in which we apply specially designed *homogeneity* transforms to each data set. We study the influence of each transform on the precision of BOF modeling for both soundscapes and music, and observe very different behaviors.

III. TEMPORAL HOMOGENEITY

A. Transform

We consider a temporal homogeneity transformation of audio data which folds an original signal onto itself a number of times (as seen in Fig. 2). The output of the twofold transform is 50%-sized random extract from the original, repeated twice. Similarly, the threefold transform is a 33%-sized extract of the original repeated three times. All signals processed by n folding from a given signal have the same duration as the original, but contain less "varied" material. Note that since the duration of the fold (an integer division of the total duration) is not a multiple of the frame duration in the general case, n folding does not simply duplicate the MFCC frames of the folded extract, but rather creates some limited jitter. The fact that all n-folded signals have the same number of frames as the original enables one to use the same modeling parameters, notably number of Gaussian components (otherwise we would have had to account for the curse of dimensionality).

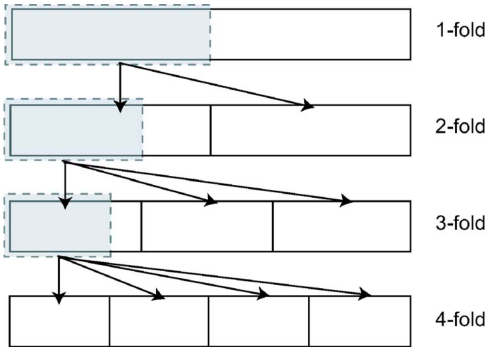

FIG. 2. (Color online) Illustration of applying three successive temporal homogeneity transforms to an audio signal, by folding it twice ("twofold"), three times ("threefold"), and four times ("fourfold"). The transform creates increasingly homogeneous signals by folding a reduced portion of the original signal. Note that the "onefold" transform is the "identity" operator.

We apply nine n-folding transforms for $n \in [1,2,3,4,5,10,20,30,50]$ to the audio signals of each data set (soundscapes and music). Each transformed signal is then processed with the above-described algorithm, namely GMM of MFCCs. This yields nine types of GMM for each original signal in a given data set, and nine similarity measures for each data set.

B. Influence on variance

Figure 3 shows the influence of n folding on the mean variance of the GMM of the transformed signals. 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.[33]

The temporal homogeneity transform has a very different influence on GMM variance when applied to urban soundscapes and music signals. The GMM variance of soundscape signals shows little dependency on temporal homogenization for ratios as low as 10% of the original signal duration. For extreme number of folds (greater than 10), the GMM variance tends to decrease slightly. This shows that the statistics of urban soundscape signals are stationary on time scales of the order of 10 s.

On the contrary, temporal homogenization has a complex influence on the GMM variance polyphonic music signals. Folding audio extracts of the original signal with durations down to 50% of the original signal's tends to reduce GMM variance. However, when the number of folds is greater than 2, the variance exponentially increases. It reaches its original 100% value when folding 15% of the signal's original duration, and increases to more than twice its original value for ratios lower than 5%. This shows that extracts smaller than half of the original duration (i.e., of the order of 100 s) are typically more heterogeneous than the overall signal in the case of polyphonic music. This indicates a rather high density of outlier frames, whose probability is overestimated when considering small extracts.

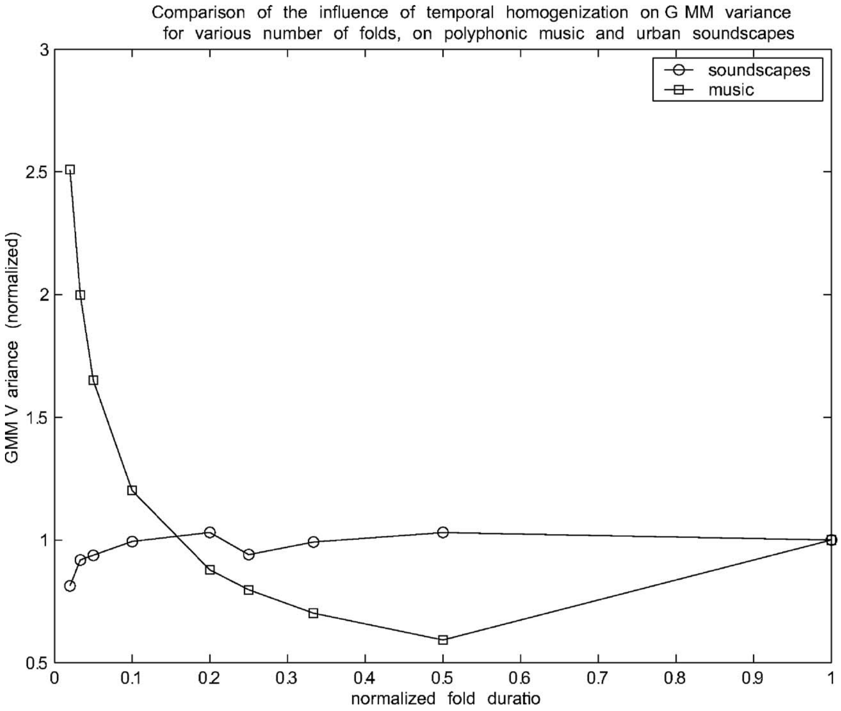

FIG. 3. Influence of temporal homogeneity transform on the mean variance of the GMMs of urban soundscapes and music signals.

C. Influence on precision

Figure 4 shows the influence of folding on the similarity of *R* precision for both classes of signals (where both precision curves are normalized with respect to their maximum). *n* folding is detrimental to the precision for both data sets. However, it appears that urban soundscapes are typically twice more robust to folding than polyphonic signals. Considering only a tenth of the audio signals cuts down precision by 15% for soundscapes, and by more than 35% for polyphonic music. In the extreme case of folding only 3 s out of a 3 min sound extract (50-folding), the precision loss is 20% for soundscapes, but more than 60% for polyphonic music.

This suggests that frame-based feature distributions for urban soundscapes are statistically much more self-similar than polyphonic music, i.e., they can be compressed to only a small fraction of their duration without much loss in terms of distribution modeling. If we authorize a 10% precision loss, soundscapes can be reduced to 10 s extracts. Polyphonic music on the contrary seems to require a large quan-

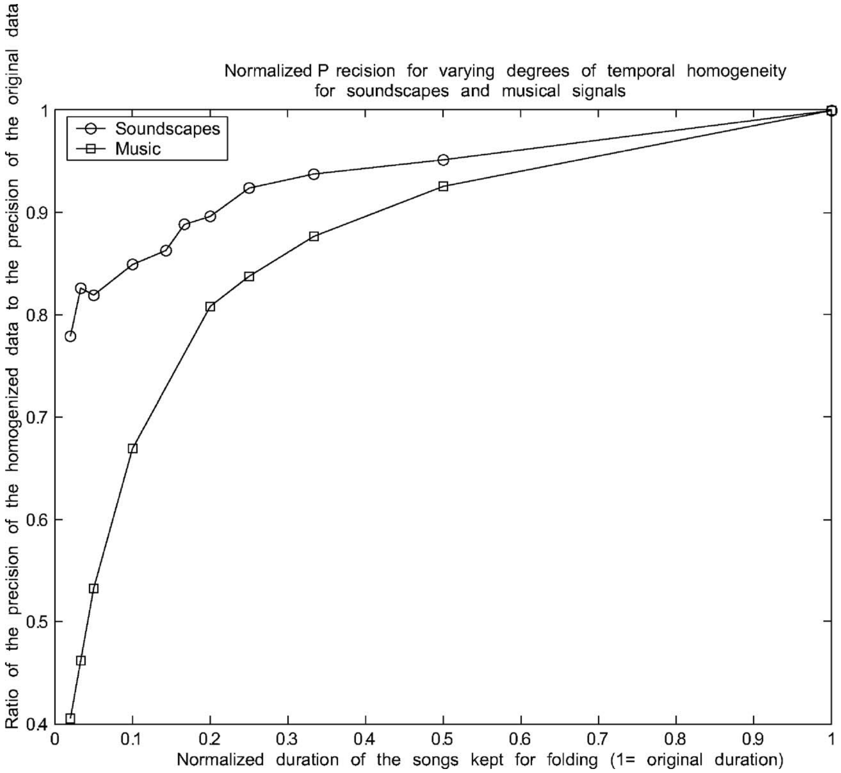

FIG. 4. Influence of temporal homogeneity transform on the precision of the similarity measure for urban soundscapes and music signals.

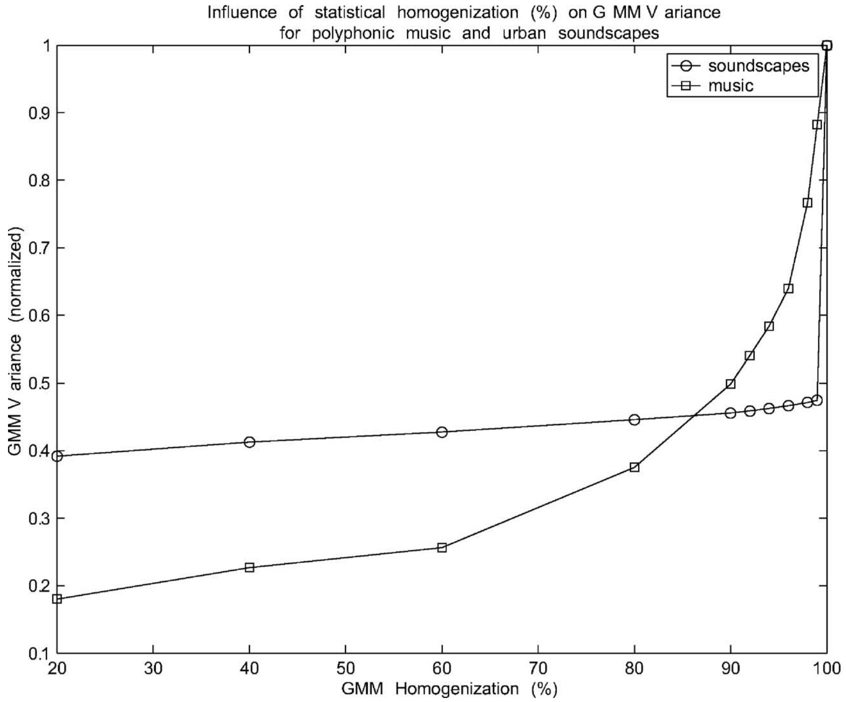

FIG. 5. Influence of statistical homogeneity transform on the variance of the GMMs of urban soundscapes and music signals.

tity of feature information in order to be properly modeled and compared: the same 10% tolerance requires more than 1 min of data.

Note that the former is comparable to the human performance[4] on the task of recognizing everyday auditory scenes (20 s). However, the latter (polyphonic music) is many times less effective than humans, who have been reported able to issue categorical judgments with good precision using as little as 200 ms of audio.[34]

IV. STATISTICAL HOMOGENEITY

A. Transform

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

```
    $g_2 = h_k(g_1)$
    $(c_1, \ldots, c_n) \leftarrow \text{sort}(\text{components}(g_1), \text{decreasing } w_c)$
    define $\mathcal{S}(i) = \sum_{j=1}^{i} weight(c_j)$
    $i_k \leftarrow \arg\min_{i \in [1,n]} \{\mathcal{S}(i) \geqslant k\}$
    $g_2 \leftarrow \text{newGMM}(i_k)$
    define $d_i = \text{component}(g_2, \text{i})$
    $d_i \leftarrow c_i, \quad \forall\, i \in [1, i_k]$
    $\text{weight}(d_i) \leftarrow \text{weight}(c_i)/\mathcal{S}(i_k), \quad \forall\, i \in [1, i_k]$
    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 does not 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 above-described similarity measure. Each transform is applied on each data set, thus yielding two sets of 11 similarity measures, the properties of which we study in the following.

B. Influence on variance

Figure 5 shows the influence of the statistical homogenization transform on the variance of the resulting GMM for both data sets. The variance of the model is evaluated with the sampling procedure already described in Sec. III B.

Again, the transformation has a very distinct influence on each type of audio signal. 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 half of the overall variance, and probably represent very different MFCC frames from the

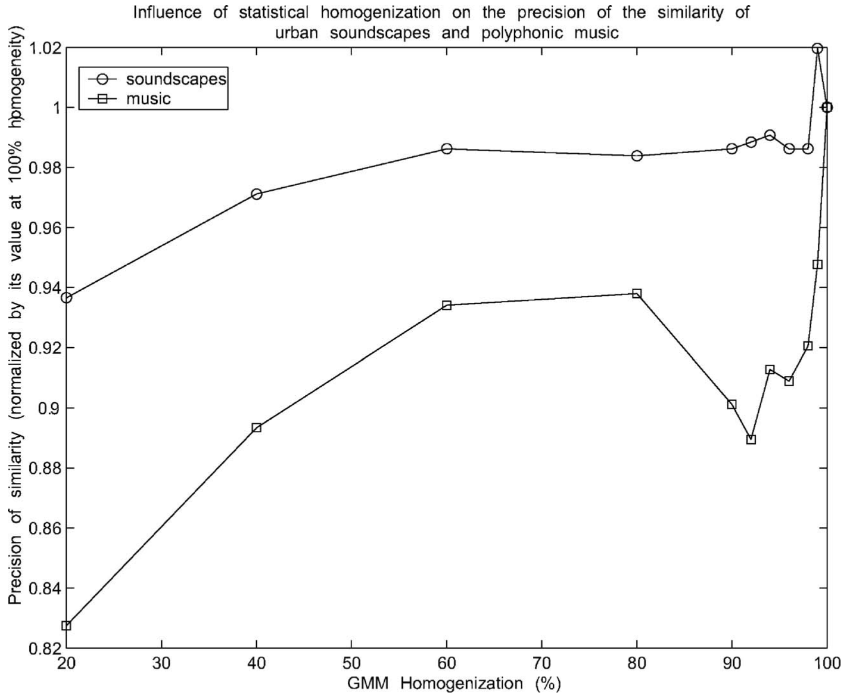

FIG. 6. Comparison of the influence of statistical homogeneity transform on the precision of the similarity measure for urban soundscapes and music signals.

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.

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.

C. Influence on precision

Figure 6 shows the influence of statistical homogenization on the precision of the resulting similarity measure, for both data sets. 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 data sets the precision is measured by reference to the baseline precision corresponding to $k=100\%$, which is different for soundscapes and music, as shown in Table III.

On both data sets, 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.

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 Fig. 5 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 6 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 modeled 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, Fig. 6 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 Fig. 6 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 modeled), it suddenly decreases when k gets higher than 60%, i.e., this new 30% information is detrimental for the modeling 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.

V. DISCUSSION

A. Physical specificities in each class of sounds

We observe critical differences in the temporal and statistical structure of the typical frame distribution for soundscapes and polyphonic music signals. The experiments reported here show that frames in polyphonic music signals are not equally discriminative/informative, and that their contribution to the precision of a simulated perceptual similarity task is not proportional to their statistical importance and long-term frequency (i.e., the corresponding component's prior probability w_c):

(1) The very informative frames for the simulation of the perception of polyphonic music (measured by their effect of acoustic similarity) are the least statistically representative (the bottom 1%).

(2) A large population of frames (in the range [60%, 95%]) is detrimental to the modeling. Another study by the authors[27] shows that the inclusion of these frames increase the hubness of a song, i.e., their statistical weight masks important and discriminative details found elsewhere in statistical minority.

Such structure cannot be observed in the frame distribution of typical urban soundscape signals.

B. A possible reason for the failure of BOF

Such differences in homogeneity for each class of signals can be proposed to explain the uneven performance of their respective modeling with the BOF approach. High performance with BOF correlates with high homogeneity: BOF-based techniques are very efficient for soundscapes, with both high precision and absence of perceptive paradoxes like hubs, while they fail for polyphonic music, which is more heterogeneous.

However, we do not give here any formal proof that heterogeneity is the main factor in explaining the failure of BOF modeling for polyphonic music signals. More complete evidence would come, e.g., by synthezing artificial signals spanning a more complete range of homogeneity values, and by comparing algorithmic predictions to human perceptive judgments.

C. Psychological relevance

The BOF approach to simulate the auditory perception of signals such as soundscapes and music makes an implicit assumption about the perceptive relevance of sound events. 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*[35] of sound events is modeled 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-mentioned model of auditory saliency would be a very crude cognitive model indeed, both to model preattentive weighting (which has been found a correlate of frequency and temporal contrasts,[36] i.e., arguably the exact opposite of statistical typicality) and higher-level cognitive processes of selective attention (which are partly under voluntary control, hence products of many factors such as context and culture[37]).

The above-presented 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 sound-scapes 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 psycho-sociological experimentations. Nevertheless, it seems at odds with a wealth of recent psychological evidence stressing that soundscapes judgment does not result in 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 versus mechanical) are systematically perceived as more pleasant.[5] 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."[38]

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. For instance, a recent study[39] in automatic instrument classification suggests that the transient part of individual notes concentrates very discriminative information for timbre identification, but that its scarsity with respect to longer steady-state information makes it difficult to exploit for machine learning algorithms. This situation of trading too little good information against too much poor-quality information is reminiscent of what we observe here. Human perception, by its higher-level cognitive processing of the structure of musical notes, gives increased saliency to frames that are otherwise in statistical minority.

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 modeled. 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.

ACKNOWLEDGMENTS

The authors would like to thank Anthony Beurivé for helping with the implementation of signal processing algorithms and database metadata management. We also thank Laurent Daudet and Pierre Leveau, as well as anonymous reviewers for their valuable comments. This research has been partially funded by the Semantic Hifi IST European project (http://shf.ircam.fr/).

[1] M. Schafer, *The Tuning of the World* (Random House, Rochester, VT, 1977).

[2] P. Lercher and B. Schulte-Forktamp, "The relevance of soundscape research to the assessment of noise annoyance at the community level," in the Eighth International Congress on Noise as Public Health Problem, Rotterdam, The Netherlands, 2003.

[3] J. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," J. Exp. Psychol. Hum. Percept. Perform. **19**, 250–267 (1993).

[4] V. Peltonen, A. Eronen, M. Parviainen, and A. Klapuri, "Recognition of everyday auditory scenes: Potentials, latencies and cues," in Proceedings of the 110th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 2001.

[5] D. Dubois, C. Guastavino, and M. Raimbault, "A cognitive approach to urban sound-scapes: Using verbal data to access everyday life auditory categories," Acta. Acust. Acust. **92**, 865–874 (2006).

[6] Note that this ("birds, children") is not intended as the definition *in intension* of a sociological representation of what a "park" is (for which we would have to give evidence and cultural context, as in Ref. 40), but only as an arbitrary example of an individual perceptual category (see Ref. 5).

[7] B. Defréville and C. Lavandier, "The contribution of sound source characteristics in the assessment of urban soundscapes," Acta. Acust. Acust. **92**, 912–921 (2006).

[8] C. Guastavino, B. Katz, J. Polack, D. Levitin, and D. Dubois, "Ecological validity of soundscape reproduction," Acta. Acust. Acust. **91**, 333–341 (2005).

[9] D. Dufournet, P. Jouenne, and A. Rozwadowski, "Automatic noise source recognition," J. Acoust. Soc. Am. **103**(5), 2950 (1998).

[10] C. Couvreur, V. Fontaine, P. Gaunard, and C. G. Mubikangiey, "Automatic classification of environmental noise events by hidden Markov models," Appl. Acoust. **54**, 187–206 (1998).

[11] M. Casey, "Mpeg-7 sound recognition tools," IEEE Trans. Circuits Syst. Video Technol. **11**, 737–747 (2001).

[12] M. Cowling and R. Sitte, "Comparison techniques for environmental sound recognition," Pattern Recogn. Lett. **24**, 2895–2907 (2003).

[13] A. Harma, J. Skowronek, and M. McKinney, "Acoustic monitoring of the patterns of activity in the office and the garden," in Proceedings of the fifth International Conference on Methods and Techniques in Behavioral Research, Wageningen, The Netherlands, 2005.

[14] B. Defréville, P. Roy, C. Rosin, and F. Pachet, "Automatic recognition of urban sound sources," in Proceedings of the 120th Audio Engineering Society Convention, Paris, France, 2006.

[15] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ, 1993).

[16] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame level noise classification in mobile environments," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Phoenix, AZ, March 1999.

[17] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP), Orlando, FL, 2002.

[18] L. Ma, D. Smith, and B. Milner, "Context awareness using environmental noise classification," in Proceedings of Eighth European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland, 2003.

[19] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surv. **34**, 1–47 (2002).

[20] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Wulton Street, Oxford, 1995).

[21] ISMIR, International Conference on Music Information Retrieval; http://www.ismir.net

[22] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," in The Second International Conference on Music Information Retrieval (ISMIR), Oct. 2001, Bloomington, Indiana.

[23] D. Liu, L. Lu, and H.-J. Zhang, "Automatic mood detection from acoustic music data," in Proceedings of the Fourth International Conference on Music Information Retrieval (ISMIR), Baltimore, MD, 2003.

[24] W.-H. Tsai and H.-M. Wang, "Towards automatic identification of singing language in popular music recordings," in Proceedings of the International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain, 2003.

[25] 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).

[26] J.-J. Aucouturier and F. Pachet, "The influence of polyphony on the dynamical modelling of musical timbre," Pattern Recogn. Lett. **28**(5), 654–661 (2007).

[27] J.-J. Aucouturier and F. Pachet, "A scale-free distribution of false positives for a large class of audio similarity measures," Pattern Recogn. http://dx.doi.org/10.1016/j.patcog.2007.04.012

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

[29] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves, a statistical analysis of speaker performance," in Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, 1998.

[30] A. Hicklin, C. Watson, and B. Ulery, "The myth of goats: How many people have fingerprints that are hard to match?," Patriot Act Report 7271, National Institute of Standards and Technology, Gaithersburg, MD (2005).

[31] F. Pachet, A. LaBurthe, A. Zils, and J.-J. Aucouturier, "Popular music access: The Sony music browser," J. Am. Soc. Inf. Sci. **55**(12), 1037–1044 (2004).

[32] B. Clarkson, N. Sawhney, and A. Pentland, "Context awareness via wearable computing," in Proceedings of the 1998 Workshop on Perceptual User Interfaces (PUIE8), San Francisco, CA, 1998.

[33] Note that a more precise measure of the width of a GMM is nontrivial to compute from the variance of its individual components (e.g., summing them, weighted with each component's prior probability), because this would have to account for the possible overlap between individual components (i.e., computing the volume of the intersection between a set of many ellipsoids in a high dimension space).

[34] D. Perrot and R. O. Gjerdinger, "Scanning the dial: An exploration of factors in the identification of musical style," in Proceedings of the 1999 Society for Music Perception and Cognition, Evanston, IL (1999).

[35] The saliency of a sound event is defined as what makes it attract auditory attention, thereby providing its weight in the representation of our envi-

ronment. Here we consider a rather permissive notion of saliency, which encompasses both preattentive mechanisms and higher-level selective attention effects.

[36] C. Kayser, C. Petkov, M. Lippert, and N. K. Logothetis, “Mechanisms for allocating auditory attention: An auditory saliency map,” Curr. Biol. **15**, 1943–1947 (2005).

[37] P. Janata, B. Tillmann, and J. Bharucha, “Listening to polyphonic music recruits domain-General attention and working memory circuits,” Cognitive, Affective and ehavioral Neuroscience **2**, 121–140 (2002).

[38] C. Guastavino, “Categorization of environmental sounds,” Can. J. Exp. Psychol., **60**(1), 54–63 (2007).

[39] S. Essid, P. Leveau, G. Richard, L. Daudet, and B. David, “On the usefulness of differentiated transient/steady-state processing in machine recognition of musical instruments,” in Proceedings of the 118th AES Convention, Barcelona, Spain, 2005.

[40] B. De Coensel and D. Botteldooren, “The quiet rural soundscape and how to characterize it,” Acta. Acust. Acust. **92**, 887–897 (2006).