



SOUNDS LIKE A PARK: A COMPUTATIONAL TECHNIQUE TO RECOGNIZE SOUNDSCAPES HOLISTICALLY, WITHOUT SOURCE IDENTIFICATION¹

PACS: 43.55.Cs

Aucouturier, Jean-Julien¹; Defreville, Boris²

1 Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan; aucouturier@gmail.com

² ORELIA, 77300 Fontainebleau, France; boris.defreville@orelia.fr

ABSTRACT

The "bag of frames" approach (BOF) to audio pattern recognition represents signals as the long-term statistical distribution of their local spectral features. This computational technique was recently used to simulate human judgements of the holistic similarity between pieces of polyphonic music ("this sounds like The Beatles"), without any direct modelling of their individual musical instruments (namely, voice, electric guitar, drums, etc.). This paper proposes to apply the same measure of acoustic similarity to natural and human sound environments (or soundscapes). We find that the approach can simulate human categorization in a simple taxonomy of urban soundscapes to near-perfect precision, better in fact than that achieved with musical signals. Such techniques to recognize environmental acoustic configurations globally without any direct modelling of their constituent sound sources can be used to disambiguate finer-but-noisier source identification algorithms: if this globally sounds like a "park", then this "car horn" must be a "bird". Based on the proposed algorithm, we discuss the difference between such contextual effects in soundscapes and music perception.

HOLISTIC VS SOURCE-BASED STRATEGIES IN AUDIO PERCEPTION

Human processing of soundscapes and polyphonic music

Psycho-physic experiments on the perception of natural and human sound environments (or soundscapes²) indicate that the cognitive processes of recognition and similarity operate mainly on the basis of the identification of the physical sources [1]. 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. This also holds for semantic categorization [2], e.g. 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). Peltonen et al. [3] reports that the average recognition time for human subjects on a list of 34 soundscapes is 20 seconds. This supports the cognitive strategy of source identification, which typically imposes long latencies, depending on the temporal density of discriminative sound events.

In contrast, while it is obviously possible to attend to sound sources in isolation in a polyphonic music stream (e.g. "listening to the piano part" in a jazz trio), humans have been reported able to issue categorical judgments of musical genres and styles with good precision using as little as 200ms of audio [4]. Such time scales indicate that sufficient information to process musical stimuli can be found in short-term, surface features of the audio signals, without requiring robust information about sound sources (instruments). Such holistic listening was also found to activate significantly different cortical areas than more analytical, source selective listening [5].

¹ 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.

² In 1977, composer R. Murray Schafer coined the term soundscape as an auditory equivalence to landscape. 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.

Computer models

There have been various attempts to simulate human perception of both soundscapes and polyphonic music with computer algorithms. Interestingly, the modelling methodologies thereto closely resemble the two alternative cognitive strategies mentioned above.

A majority of algorithms modelling *soundscape* perception takes the strategy to identify the constituent sound sources individually, as humans do [6-8]. Typical implementations describe sound extracts with generic frame-level features, such as MPEG-7 spectral descriptors, and use hidden Markov models [10] to represent their statistical dynamics. Recent research [8] proposes to enhance this typical scheme by learning problem-specific features, adapted to each sound class, with genetic programming.

In contrast, most algorithms concerned with recognition and similarity of *polyphonic music* directly model music as a whole, without prior identification of constituent sound sources (instruments). In these works (see [9] for a review), music is 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” (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 (e.g. 50ms with a 50% overlap), and for each frame, a feature vector is computed. Features usually consists of a generic, all-purpose spectral representation such as Mel Frequency cepstrum Coefficients (MFCC, [10]). 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 [11]) which models the global distributions of the features of signals corresponding to each class (e.g. “rock music”). 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.

A model for holistic soundscape perception

This paper reports on experiments to simulate human processing of soundscapes, using the holistic strategy typically employed for musical audio signals. We apply to a dataset of urban soundscapes an algorithmic measure of acoustic similarity that we introduced in the context of polyphonic music [9]. 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”. We find here that the measure is nearly optimal for modelling the perceptual similarity of urban soundscapes, better in fact than when applied to musical signals. We propose that such techniques to recognize environmental acoustic configurations globally without any direct modelling of their constituent sound sources can be used to remove ambiguities occurring with finer-but-noisier source identification algorithms.

ALGORITHM AND DATASETS

Algorithm

We sum up here the audio similarity algorithm presented in [9]. 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 $\log S$.

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

(1)

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:

$$p(x_t) = \sum_{m=1}^{m=\mathcal{M}} \pi_m \mathcal{N}(x_t, \mu_m, \Sigma_m) \quad (2)$$

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 [11].

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 in (1)) is defined by:

$$d(A, B) = \int p_A(x) \log \frac{p_B(x)}{p_A(x)} dx \quad (3)$$

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

$$\widetilde{d(A, B)} = \frac{1}{n} \sum_{i=1}^n \log \frac{p_B(x_i)}{p_A(x_i)} \quad (4)$$

(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, namely 20 MFCCs appended with 0th order coefficient, 50-component GMMs, compared with $n = 2000$ Monte-Carlo draws.

Datasets

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. Polyphonic Music

The polyphonic music dataset used in this study contains 350 popular music titles, extracted from the Cuidado database. 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 [9].

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 than the seed. This is identical to the methodology used e.g. in [9].

RESULTS

Precision

Table III gives the precision of timbre similarity applied to both datasets. It appears that the results are substantially better for urban soundscapes than for polyphonic music signals, nearing perfect precision in the first 5 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).

Table 2: Comparison of similarity measure precision 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

Hubs

Careful analysis shows that the application of the similarity measure to polyphonic music signals 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 [12]). However, it appears that such hub items are not found when applying the technique to soundscapes signals.

A natural measure of the “hubness” of a given item (song or soundscape) is the number of times the item occurs in the first n nearest neighbors of all the other items in the database (so called “number of n -occurrences”). Note that the mean n -occurrence of a nitem is equal to n , independently of the database and the distance measure.

Figure 1 shows the histogram of the number of 20-occurrences N_{20} obtained with the above 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 dataset than in the urban soundscape dataset. This is also confirmed by the manual examination of the similarity results for the urban soundscapes: none of the (few) false positives re-occur 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, among whom polyphonic music, but not urban soundscapes.

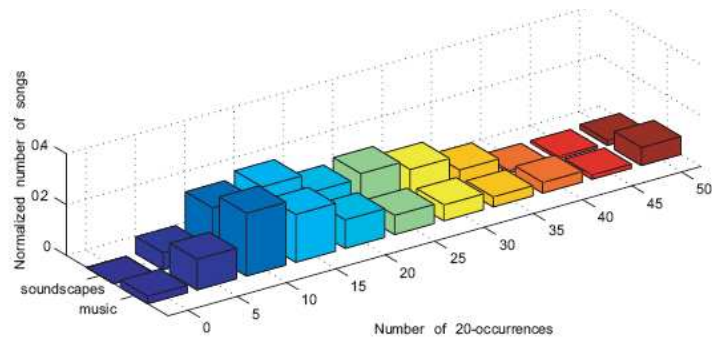


Figure 1: Comparison of the histograms of number of 20-occurrence for the same measure used on urban soundscapes and polyphonic music

DISCUSSION: CONTEXT EFFECTS

Disambiguating source identification

Our study shows that BOF pattern recognition techniques are able to recognize environmental acoustic configurations with near-perfect precision, and this without any direct modelling of their constituent sound sources. Previous research shows that the performance of typical source identification algorithms for soundscape signals depends a lot on the composition of the test dataset. As seen in Table 3 (reproduced from [8]), it is easier to computationally discriminate bus sounds against other vehicles sounds than against more general corpuses including bird twitter and human voice (see [8] for details). This suggests that BOF techniques can be used as a pre-selection stage to filter out candidate sound sources that are unlikely in a given acoustic environment: if the whole scene globally sounds like a "boulevard", then this "bird" must be a "car horn".

Table 3: Dataset effects in computer source recognition for urban soundscapes (reproduced from [8]). “EDS, kNN” denotes k-nearest neighbor classification based on

genetic-algorithm based signal features, while “MFCC, GMM” denotes techniques similar to that explored in this paper (see original paper for details)

Classification Task	Test Precision (%) (EDS, kNN)	Test Precision (%) (MFCC, GMM)
“Bus” vs “Other vehicles” (car, truck, moped, motorcycle)	85.2	79.4
“Bus” vs “All” (as above + bird, voice)	67.0	69.1

More context in music than soundscapes ?

While a computer simulation of such contextual inference would likely improve computer recognition, our current findings question the importance of context in the human processing of soundscapes.

The BOF approach to auditory perception builds an amorphous and holistic description of the object being modeled. The fact that such a simple model is sufficient to simulate the perception of soundscapes, but not of polyphonic music, could suggest that the cognitive processes involved in the human processing of the former are less “demanding” than for the latter. This finding 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 [2].

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 (which is what BOF does, in a sense) hurts the perception of music more than that of soundscapes. This naturally will have to be validated with proper psycho-sociological experimentations.

References:

- [1] J. Ballas, “Common factors in the identification of an assortment of brief everyday sounds”, *Journal of Experimental Psychology, Human Perception and Performance* 19, 250–267 (1993).
- [2] 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).
- [3] V. Peltonen, A. Eronen, M. Parviainen, and A. Klapuri, “Recognition of everyday auditory scenes: potentials, latencies and cues”, in *Proc. 110th Convention of the Audio Engineering Society*, 2001.
- [4] D. Perrott and R. Gjerdingen (1999). *Scanning the dial : an exploration of factors in the identification of musical style*. In *Proceedings of the 1999 Society for Music Perception and Cognition*.
- [5] P. Janata, B. Tillmann and J. Bharucha, “Listening to polyphonic music recruits domain-general attention and working memory circuits”, *J. Cognitive, Affective & Behavioral Neuroscience* 2002, 2 (2), 121-140
- [6] D. Dufournet, P. Jouenne, and A. Rozwadowski, “Automatic noise source recognition”, *Journal of the Acoustical Society of America* 103, 2950 (1998).
- [7] P. Gaunard, C. G. Mubikangiey, C. Couvreur, and V. Fontaine, “Automatic classification of environmental noise events by hidden markov models”, *Applied Acoustics* (1998).
- [8] B. Defreville and C. Lavandier, “The contribution of sound source characteristics in the assessment of urban sound-scapes”, *Acta Acustica united with Acustica* 92, 912–921 (2006).
- [9] 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).
- [10] L. Rabiner and B. Juang, *Fundamentals of speech recognition* (Prentice-Hall) (1993).
- [11] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford Press) (1995).
- [12] 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).