Content-based Retrieval of Environmental Sounds by Multiresolution Analysis

Ianis Lallemand
UMR STMS
IRCAM–CNRS–UPMC
Paris, France
ianis.lallemand@ircam.fr

Diemo Schwarz
UMR STMS
IRCAM–CNRS–UPMC
Paris, France
diemo.schwarz@ircam.fr

Thierry Artières
LIP6
CNRS–UPMC
Paris, France
thierry.artieres@lip6.fr

ABSTRACT

Query by example retrieval of environmental sound recordings is a research area with applications to sound design, music composition and automatic suggestion of metadata for the labeling of sound databases. Retrieval problems are usually composed of successive feature extraction (FE) and similarity measurement (SM) steps, in which a set of extracted features encoding important properties of the sound recordings are used to compute the distance between elements in the database. Previous research has pointed out that successful features in the domains of speech and music, like MFCCs, might fail at describing environmental sounds, which have intrinsic variability and noisy characteristics. We present a set of novel multiresolution features obtained by modeling the distribution of wavelet subband coefficients with generalized Gaussian densities (GGDs). We define the similarity measure in terms of the Kullback-Leibler divergence between GGDs. Experimental results on a database of 1020 environmental sound recordings show that our approach always outperforms a method based on traditional MFCC features and Euclidean distance, improving retrieval rates from 51% to 62%.

1. INTRODUCTION

1.1 Background

Ever-larger sound databases, nowadays easily available on the Internet or in sound banks, represent a great potential for musical creativity. Yet most of this potential may remain hidden to the user without efficient means of exploring the data: in 2005, for instance, the total duration of the Creative Commons sound archive *archive.org* was estimated between 34.2 and 2,000 years [**?**], and has been increasing ever since. Consequently, methods focusing on content-based sound information retrieval, like query by example (QBE), have attracted much attention in the past years, as they allow easier and faster access to relevant soundfiles than usual keywords-based search.

In this work, we focus on environmental sound recordings, which are prominent in most commercial databases as the latter are mainly aimed at sound designers working in the film, game and virtual reality industries. Environmental sounds are sought after by composers and artists as well, who use them either as *concrete* sound objects or as raw materials in granular or concatenative synthesis techniques [**?**]. Our core application is a QBE retrieval system, to which the user provides a short example (typically 2 to 5 seconds) of the kind of environmental sound he is interested in. After analysis of the acoustic and structural properties of the sound, the system searches the database for a given number of relevant recordings. By making manual specification of keywords unnecessary, such systems allow users to retrieve sounds on the basis of their sonic properties alone: this approach gives access to a greater variety of relevant sounds than a keyword-based system, which can be biased by the way recordings have been tagged in the database.

Other applications include metadata suggestion to users submitting recordings to online databases like *Freesound* [1] . The system could suggest tags for their submission by retrieving relevant database recordings, then accessing the metadata they have been given by previous users. This could help to harmonize the way similar sounds are labelled, making keyword-based searches easier. Note that QBE is not incompatible with keyword search schemes, which can always be used to filter the system's output.

1.2 Problem

QBE systems are based on a notion of similarity between sound recordings. Defining similarity first involves a feature extraction (FE) step, during which a set of features that precisely represent a recording are computed. Features are a much shorter description of the soundfile than raw signal data; like the widely used Mel-frequency cepstral coefficients (MFCCs), most features are based on time-frequency representations such as the short-term Fourier transform (STFT) or the wavelet transform (DWT). The extracted features are then used for similarity measurement (SM), which consists in computing the distance between the query recording and each database image to return the N closest matches.

By definition, environmental sounds present a greater variability than music or speech. This makes the feature extraction step difficult in the sense that one has to look for features that can describe a wide range of sounds, without assumptions on their timbral properties or structure. While

[1] http://www.freesound.org

MFCCs are widely used with excellent results for structured signals like speech and music, their performances degrade in the presence of noise, like found in non-clearly structured environmental sounds. Noisy signals that have a flat spectrum, like rain and insect chirping, may also be inefficiently analyzed by MFCCs [**?**].

Another difficulty is that environmental sounds contain details at very different time scales, that cannot be captured by fixed-size sliding-window techniques like STFT. These problems motivate the development of a multiresolution analysis tool capable of describing a wide class of signals, without assumptions on the nature of events happening within them.

1.3 Our Approach

We present in section 3 a multiresolution feature extraction method based on the modeling of wavelet subband coefficients' distributions with generalized Gaussian distributions (GGDs). We use the symmetrized Kullback-Leibler divergence (KLD) as similarity measure between two sound recordings. This method extends the approach developed by Do and Vetterli [**?**] for content-based retrieval of texture images to one-dimensional audio signals. Our main contribution is a QBE retrieval system for environmental sounds, in which the similarity between the query and the database elements is computed using GGD wavelet features with KLD. We present and evaluate this retrieval system in section 4.

As detailed in section 4.3, it occurred to us that no reference dataset has emerged in the domain of environmental sound recognition. Thus, to evaluate our retrieval system, we collected and labelled a large number of environmental sounds recordings from the widely-used *Sound Ideas Series 6000* library. We provide details about this set in section 4.3, as well as an Internet link to a list of filenames allowing one to build it back from the *Sound Ideas* library.

2. RELATED WORK

2.1 Environmental Sound Recognition

Compared to the speech and music domains, the field of environmental sound recognition has few publications. While most works point out the limits of conventional features such as MFCCs [**?**, **?**], only a few of them present novel features [**?**]. More importantly, few works investigate the use of similarity measures [**?**, **?**, **?**]. This is because most works address the problem of classifying environmental sounds in a given set of classes, which are known *a priori* [**?**, **?**, **?**]. Hence the "closeness" of sounds is determined by a classifier (Gaussian mixture model, k-nearest neighbor, etc.) which has some knowledge about the distribution of the sounds amongst classes. On the contrary, a QBE retrieval system doesn't assume the existence of classes, and is only concerned by returning N relevant sounds when given a specific query. Hence defining a similarity measure between two sounds is essential in this context.

Xue *et al.* [**?**, **?**] use a switching state-space model (SSM) to measure the similarity between a given query and the database sounds, using common perceptual features (RMS level, Bark weighted spectral centroid, etc.). A comparison of existing features and distance measures performed by Cowling and Sitte [**?**] show that wavelet-based features tend to give good recognition results, combined with dynamic time warping (DTW). However, while this motivates the use of wavelet-based features, the high computational cost of DTW invites to search for other similarity measures.

2.2 Environmental Sound Synthesis

Most environmental sound synthesis works are based on the modeling of the statistical properties of textures [**?**], which are studied on extracted features. Although our aim is a QBE retrieval system and not sound synthesis, such works provide valuable information about which features seem successful at encoding the variability of environmental sounds.

The feature extraction method detailed in section 3 is essentially based on the modeling of wavelet subband statistics with GGDs. Our assumption is that these statistics encode characteristic features of environmental sounds. It is supported by works like that of McDermott *et al.* [**?**], in which the authors synthesize stationary environmental sounds by matching the statistics of a sample of noise with those of real sounds. Furthermore, Dubnov *et al.* [**?**] and O'Regan and Kokaram [**?**] use wavelet subband coefficients to model the statistics of environmental sounds (stationary and non stationary), yielding state of the art synthesis results. This suggests that wavelet-based features might perform well with environmental sounds.

2.3 Texture Image Recognition

The field of texture image recognition has more publications than its sound counterpart. The wavelet transform is used in various feature extraction approaches for retrieval applications. The core idea is that the energy in wavelet subbands can be used to identify a texture. Wouwer *et al.* [**?**] proposed to extend energy-based methods by adopting a more precise model of subband coefficients. Instead of considering the subband's energy alone, they use generalized Gaussian distributions to fit the subband coefficients' histogram. This approach was extended by Do and Vetterli [**?**], who derived a closed-form version of the Kullback-Leibler divergence between two GGDs.

3. WAVELET-BASED FEATURES AND SIMILARITY MEASURE

3.1 Wavelet Representation

The wavelet transform is a multiresolution, time-frequency signal representation [**?**]. Each level encodes the signal's information at a particular resolution, and is non-redundant with the next smaller resolution levels. The wavelet transform of a one-dimensional signal can be computed using a cascade filter bank (the so-called "pyramid" architecture), as shown in figure 1. The signal S_0 is first split into a low-pass (or scaling) series of coefficients S_{-1} by convolving

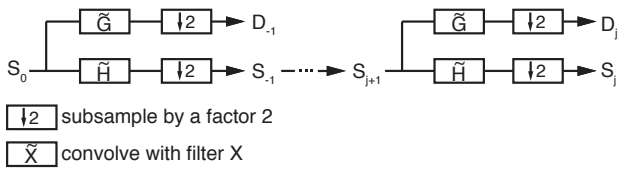

Figure 1: Pyramid architecture for computing the wavelet transform of a one-dimensional signal.

the original signal with a low-pass filter $\tilde{H}$, and subsampling by a factor of 2. In parallel, the series of wavelet (or detail) coefficients D_{-1} is computed by convolving the signal with a wavelet filter $\tilde{G}$, and subsampling by a factor of 2. The same filters can be applied again on the scaling coefficients S_{-1} to obtain coefficients S_{-2} and D_{-2}, and so forth.

There is the same number of series of detail coefficients as there are levels (referred to as *subbands*) in the wavelet transform. A three-level wavelet transform would give series of detail coefficients D_{-1}, D_{-2} and D_{-3}, and the series of scaling coefficients S_{-3} which correspond to the three-level approximation of the signal. Because each filter response is subsampled by a factor 2, an N-level wavelet transform requires a signal S_0 of dyadic length 2^J, with $J \geq N$. The series of detail coefficients at level i (S_{-i}) would then be a signal of length 2^{J-i}.

In our experiments, we used the Daubechies maximally flat orthogonal filters of length 8 (D_4) as wavelet filters. Evaluation of all families of wavelet filters is beyond the scope of this paper; however, we should note that the Daubechies filters are a common choice in both image and sound processing domains [**?**, **?**].

Sections 3.2 and 3.3 present the theoretical framework from which the features and the similarity measure used in our retrieval system are derived. It is based on the work of Do and Vetterli [**?**], who originally applied it to describe two-dimensional texture images.

3.2 Modeling of Subband Coefficients by GGDs

We model the distribution of each subband's series of detail coefficients with a generalized Gaussian density. We discard the lowest band scaling (approximation) coefficients as they do not encode sailent details at the chosen decomposition depth. Generalized Gaussian distributions are defined as:

$$p\left(x;\alpha,\beta\right) = \frac{\beta}{2\alpha\Gamma\left(1/\beta\right)} e^{-\left(|x|/\alpha\right)^{\beta}}, \tag{1}$$

where Γ is the Gamma function. For $x > 0$:

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1} dt. \tag{2}$$

They can be seen as an extension of Gaussian and Laplacian distributions, which correspond to $\beta = 2$ and $\beta = 1$. Parameter α (*scale* parameter) models the width of the distribution (standard deviation), while β (*shape* parameter) is inversely proportional to the decreasing rate of the distribution's tail. Given the series of detail coefficients $D_{-j} = (D^1_{-j}, ..., D^n_{-j})$ for subband j, a $\hat{\beta}$ estimation can be obtained by:

$$\begin{aligned} 1 + \frac{\Psi\left(1/\hat{\beta}\right)}{\hat{\beta}} - \frac{\sum \left|D^p_{-j}\right|^{\hat{\beta}} \ln \left|D^p_{-j}\right|}{\sum \left|D^p_{-j}\right|^{\hat{\beta}}} \\ + \frac{\ln\left(\frac{\hat{\beta}}{n}\sum \left|D^p_{-j}\right|^{\hat{\beta}}\right)}{\hat{\beta}} = 0, \end{aligned} \tag{3}$$

where all sums are taken from $p = 1$ to n, and Ψ is the digamma function, i.e. $\Psi(x) = \Gamma'(x)/\Gamma(x)$. A $\hat{\alpha}$ estimation is given by:

$$\hat{\alpha} = \left(\frac{\hat{\beta}}{L}\sum_{p=1}^{n} \left|D^p_{-j}\right|^{\hat{\beta}}\right)^{1/\hat{\beta}}. \tag{4}$$

Equation 3 is *transcendental*. We solve it numerically using the Newton-Raphson iterative procedure [**?**]. The estimated $\hat{\beta}$ parameter can then be used to obtain $\hat{\alpha}$ with equation 4. Figure 2 shows two histograms of wavelet subband coefficients, and the GGD fits. Note that, if the first histogram could be modeled by a standard Gaussian distribution ($\hat{\beta}$ close to 2), the second histogram clearly shows the benefits of the use of GGD fits, as a standard Gaussian fit would fail at producing such a peaked distribution (we find $\hat{\beta}$ = 1.5281 by fitting with a GGD).

3.3 Distance Measure Between GGDs

The Kullback-Leibler divergence, or relative entropy, provides a statistical, non-symmetric measure of the distance between two probability density functions (PDFs) [**?**]. The KLD between PDFs $p\left(.;\boldsymbol{\theta}_q\right)$ and $p\left(.;\boldsymbol{\theta}_i\right)$ is:

$$D\left(p\left(.;\boldsymbol{\theta}_q\right)||p\left(.;\boldsymbol{\theta}_i\right)\right) = \int p\left(x;\boldsymbol{\theta}_q\right) \ln \frac{p(x;\boldsymbol{\theta}_q)}{p(x;\boldsymbol{\theta}_i)} dx. \tag{5}$$

Although this is usually not the case for most PDFs, a closed-form version of the KLD between two GGDs can be found [**?**]:

$$\begin{aligned} D\left(p\left(x;\alpha_1,\beta_1\right)||p\left(x;\alpha_2,\beta_2\right)\right) = -\frac{1}{\beta_1} \\ + \ln\left(\frac{\beta_1\alpha_2\Gamma\left(1/\beta_2\right)}{\beta_2\alpha_1\Gamma\left(1/\beta_1\right)}\right) \\ + \left(\frac{\alpha_1}{\alpha_2}\right)^{\beta_2} \frac{\Gamma\left(\left(\beta_2+1\right)/\beta_1\right)}{\Gamma\left(1/\beta_1\right)}. \end{aligned} \tag{6}$$

This allows us to compute a similarity measure between two different sound recordings, at a given subband. Besides being theoretically motivated, the use of KLD as a similarity measure provides a closed-form formula, yielding lighter computational cost than DTW for instance. This measure can be computed in terms of the two GGDs' parameters, which also significantly reduces the space needed to store the recordings' features.

To define the overall similarity measure between two recordings (taking into account all subbands), we make

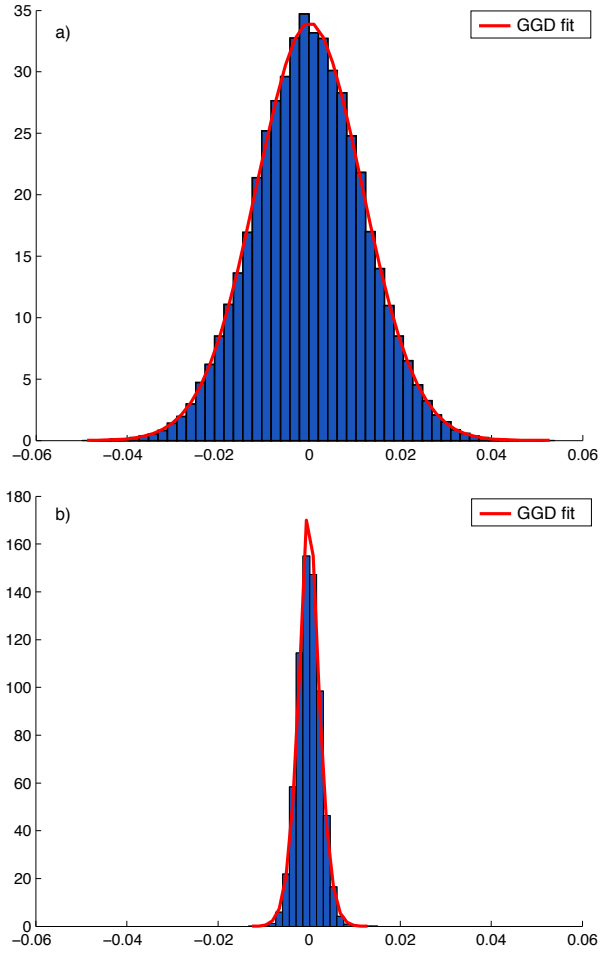

Figure 2: GGD fits of two histograms of second wavelet subband coefficients. Two different subclips extracted from the same recording of ambient city sounds were used. *a)* subclip 1. The estimated parameters are $\hat{\alpha} = 0.0166$ and $\hat{\beta} = 1.9462$.
b) subclip 2. The estimated parameters are $\hat{\alpha} = 0.0030$ and $\hat{\beta} = 1.5281$.

the simplifying assumption that these subbands are statistically independent. As shown by Dubnov *et al.* [**?**], there may in fact exist some correlations between subbands. However, we assume that our independent-subband model will have enough discriminating power amongst the environmental sound class to be used in a retrieval context. Hence we define the distance between two sound recordings R_1 and R_2 as the sum of the symmetrized KLDs between each subband pair:

$$D(R_1, R_2) = \sum_{i=1}^{N} D\left(p\left(x;\alpha_1^i,\beta_1^i\right)||p\left(x;\alpha_2^i,\beta_2^i\right)\right) + D\left(p\left(x;\alpha_2^i,\beta_2^i\right)||p\left(x;\alpha_1^i,\beta_1^i\right)\right), \quad (7)$$

Where α_k^i and β_k^i are the GGD parameters from the wavelet subband i of the recording R_k, and N is the number of levels (subbands) in the wavelet decomposition.

4. EVALUATION: QBE RETRIEVAL SYSTEM

4.1 Retrieval System

We evaluate our similarity measure in a QBE retrieval application. Given a query i, our system computes the distance between i and each element of a database composed of environmental sound recordings. The distances are then sorted in ascending order to allow the system to return the indices of the N closest (or most relevant) recordings.

The expected output of a retrieval system is a series of sounds relevant to a user's query. To evaluate our system, we use a standard precision-recall procedure. For each query sound clip i, we retrieve the N most relevant sound clips, then define precision "per sound clip" $P(i)$ as:

$$P(i) = \frac{m_i}{N}. \quad (8)$$

Where $m_i \leq N$ is the number of relevant sound clips amongst the N retrieved sound clips. Hence precision is defined as the number of relevant retrieved sound clips over the number of retrieved sound clips. To define recall "per sound clip" $R(i)$, we assume that there exist L sound clips relevant to query i in all our dataset; for simplicity of exposition, we assume that L is the same for each query. Recall is then defined as the number of relevant retrieved sound clips over the *total* number of relevant sound clips:

$$R(i) = \frac{m_i}{L}. \quad (9)$$

The quantities $P(i)$ and $R(i)$ are averaged on all possible queries to obtain "global" precision $\bar{P}$ and recall $\bar{R}$. The process is then iterated by varying the number N of retrieved sound clips from 1 to N_{max} (usually $N_{\text{max}} \gg L$ to ensure that all relevant sound clips have a chance to be retrieved).

The F-measure "per sound clip" $F(i)$ is defined as the harmonic mean of $P(i)$ and $R(i)$:

$$F(i) = 2\frac{P(i)R(i)}{P(i)+R(i)}. \quad (10)$$

The overall performance of a retrieval system can thus be evaluated using the "global" F-measure $\bar{F}$, obtained by averaging $F(i)$ on all possible queries. As precision and recall, the F-measure is a number located between 0 and 1, 1 being the best value.

4.2 Evaluation Methodology

To define the L relevant sound clips for each query, we adopt a standard methodology in the field of texture image recognition [**?**, **?**, **?**]. The basic idea is to use homogeneously-sounding source recordings, which can be split into series of sound clips. By *homogeneously-sounding*, we refer to recordings in which the number and nature of sound sources remain constant throughout their whole duration. For instance, we didn't allow a recording of machine noises to include isolated conversation noises, in a similar spirit to widely-used texture images datasets like the MIT Vision Texture database [2] (VisTex). In this

[2] http://vismod.media.mit.edu/vismod/imagery/VisionTexture/

context, two sound clips extracted from the same file can be defined as perceptually relevant to each other. This methodology provides a convenient way of performing a first evaluation of our retrieval system, since no manual definition of relevance (e.g. manual labeling of each sound clip) is needed. We acknowledge however that it should be complemented with a human-based evaluation, which we plan to perform in the future.

4.3 Dataset

Most reference databases in the speech and music recognition communities are freely available, providing a convenient way of evaluating results on the same test sets as other publications. To our knowledge, no comparable evaluation set has emerged for applications related to environmental sounds. As a consequence, we decided to collect a large number of environmental sounds from the *Sound Ideas Series 6000* commercial library [3]. This 40-CD collection is mainly aimed at sound designers working in the film industry, which makes it a relevant choice for our retrieval application. It is composed of short sound effects and longer ambience recordings.

The selection process involved discarding the sound effects and very short recordings, as well as environmental sound clips in which precise identification of sound sources was impossible. The remaining recordings form a small subset of the original *Sound Ideas* database (358 files over 7546, or about 4.7 %). They cover a broad spectrum of environmental sound types (e.g. *bubbles*, *conversation*, etc.). Class labels (i.e. sound source types) were attributed manually for each sound, as such information was not given by the *Sound Ideas* library [4]. A list of filenames allowing one to build our database back from the *Sound Ideas* library is available online [5]. Manual attribution of class labels as well as pre-selection of homogeneously-sounding recordings make it possible to use this dataset in classification problems without further labeling work.

Table 1 shows the number of sound clips per type of sound source used in our evaluation dataset. To comply with our evaluation methodology requiring the same number of sound clips to be extracted from each source recording, we select the recordings which are lengthy enough amongst those in our *Sound Ideas*-based dataset, i.e. 68 recordings. We split each of them in 15 monophonic sound clips of length 2^{18} samples, or about 5 s at 44100 Hz. We then iteratively select each of the 1020 obtained sound clips as a query. For each query, we use the 1019 remaining sound clips as a retrieval database, inside which we define the query's relevant sound clips as the 14 other sound clips that have been extracted from the same recording as the query (hence $L = 14$).

4.4 Results

Table 2 shows the results obtained by our system when retrieving $N = 1$ up to $N = L = 14$ top matches. The minimum and maximum values of $\bar{P}$ and $\bar{R}$ are given as *Min. precision*, *Max. precision*, *Min. recall* and *Max. recall*. The entry *Recall (N=14)* shows the recall (or retrieval rate) obtained when retrieving 14 top matches. We compare our method, which uses GGD-based features and KLD, with a method based on MFCC features and normalized Euclidean distance (ED). To outline the interest of our KLD-based similarity measure, we have also included the results of a method based on GGD features and Euclidean distance. This allows to compare the gain in using GGD features over MFCC features with the same Euclidean similarity measure. Note that a similar case using the KLD similarity measure can't be included, as the KLD only makes sense as a measure of the distance between two PDFs. To use it with MFCCs would require building an underlying probabilistic model for MFCC features, which is beyond the scope of this paper. This further underlies the interest of our feature extraction scheme based on GGD modeling of wavelet subbands, which naturally enables the use of the KLD as a similarity measure.

Class	#soundfiles
ambience (city)	30
ambience (crickets)	30
ambience (crowd)	195
ambience (crowd with children)	45
ambience (forest)	75
ambience (machine)	60
ambience (plane)	30
ambience (riot)	30
ambience(shopping mall)	60
ambience (subway)	30
bubbles	30
fire	30
conversation	60
jet engine	15
race	45
water (flowing)	225
wind	30
Total	1020

Table 1: Number of sound clips per sound source type in evaluation dataset.

GGD-based features were computed using a six-level wavelet decomposition. We computed the MFCCs on *Matlab* with the *Auditory Toolbox* [**?**], using 13 coefficients, a window size of 256 samples and a framerate of 100 Hz. We found that the inclusion of first-order derivatives amongst MFCC features and removal of first MFCC coefficient yielded lower retrieval rates. As a consequence, we present the results obtained with MFCCs alone.

Figure 3 shows the precision-recall curves ($\bar{P}$ vs. $\bar{R}$) obtained with the GGD+KLD and MFCC+ED methods, as well as plots of the F-measure against recall ($\bar{F}$ vs. $\bar{R}$) [6].

[3] http://www.sound-ideas.com/6000.html

[4] Note that our QBE retrieval system application, in which no sound classes are known *a priori*, doesn't make use of such information.

[5] http://imtr.ircam.fr/imtr/Environmental_Sound_Dataset

[6] Notice that the 14th values of recall and F-measure are equal. This is because by definition, $\bar{P} = \bar{R}$ for $N = L = 14$.

	MFCC+ED	GGD+ED	GGD+KLD
Min. precision	0.5131	0.5586	**0.6174**
Min. recall	0.0482	0.0553	**0.0637**
Min. F-meas.	0.0899	0.1031	**0.1188**
Max. precision	0.6745	0.7735	**0.8912**
Max. recall	0.5131	0.5586	**0.6174**
Max. F-meas.	0.5131	0.5586	**0.6174**
Recall (N=14)	0.5131	0.5586	**0.6174**

Table 2: Minimum and maximum values of evaluation results obtained when retrieving $N = 1$ up to $N = L = 14$ matches. Recall (or retrieval rate) values are given for $N = 14$. Best results are displayed in bold font.

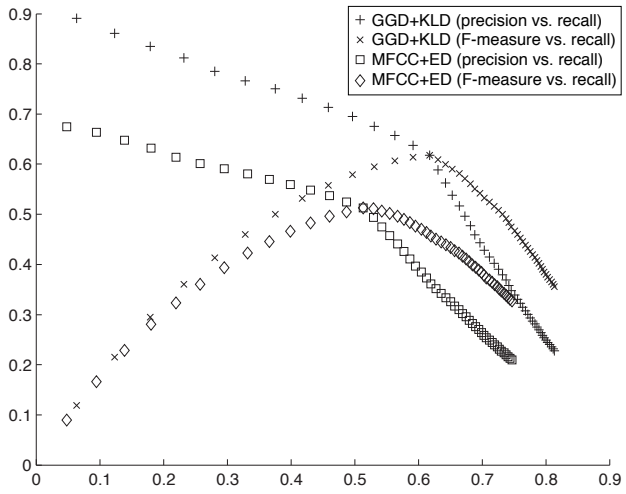

Figure 3: Precision-recall curves obtained with the GGD+KLD and MFCC+ED retrieval methods when retrieving from $N = 1$ up to $N = 50 > L$ matches. Also shown are plots of F-measure against recall.

Each point corresponds to the results obtained when retrieving from $N = 1$ up to $N = 50 > L$ top matches. The minimum and maximum values of the curves within the 14 first points' range correspond to those of table 2.

We observe in figure 3 that the precision-recall curve of the GGD+KLD method is always above that of MFCC+ED, which is also the case for the plot of F-measure against recall. Hence our method (GGD+KLD) always outperforms the method based on traditional features and similarity measure (MFCC+ED). As shown in table 2, our retrieval system is able to retrieve about 62% of relevant sound clips in the 14 top matches, whereas the traditional approach based on MFCCs and Euclidean distance only achieves a 51% retrieval rate.

Table 2 shows that the method based on wavelet features and Euclidean distance (GGD+ED) also outperforms the MFCC+ED approach. As these two retrieval systems only differ in the choice of features, the notably superior results of the GGD+ED method shows that the GGD wavelet features provide a better description of environmental sounds properties than MFCCs. This confirms our assumption that our wavelet-based features encode important informations about the characteristics of environmental sounds. As expected, the best results are obtained when using the wavelet features in conjunction with the KLD, i.e. when combining informative features with a proper, statistically-motivated similarity measure.

5. CONCLUSION

We have introduced a novel set of multiresolution features as well as a consistent similarity measure for QBE retrieval of environmental sounds. Our approach consists in modeling the distribution of wavelet subband coefficients by a generalized Gaussian density. Assuming that subbands are statistically independent, we propose a closed-form version of the overall distance between two environmental sound recordings using the symmetrized Kullback-Leibler divergence.

We have collected a database of 1020 sound clips from the *Sound Ideas* library, originating from 68 different source recordings covering a broad spectrum of environmental sound types. Experimental results on this database show that our approach always outperforms a method based on traditional MFCC features and Euclidean distance, improving retrieval rates from 51% to 62%. Results also show that the proposed method benefits simultaneously from its set of features and its proper similarity measure.

Further improvements on our approach could be made by considering multi-dimensional generalized Gaussian density distributions to model the joint probability distribution of wavelet subband coefficients. Another application would be to use our method in classification-related domains which have few publications devoted to environmental sounds, like computational auditory scene recognition (CASR). We are currently investigating the use of support vector machines with KLD-based kernels to apply our method to classification problems.

Finally, we plan to distribute a *Freesound*-based dataset of environmental sound recordings. The building of a Creative Commons-licensed evaluation bank, though demanding more time than for commercial libraries-based datasets, is likely to allow more applications to benefit from it. In particular, we will be conducting a user-based study of the performances of our retrieval system on a collection of *Freesound* recordings.

6. ACKNOWLEDGMENTS

The work presented here is partially funded by the *Agence Nationale de la Recherche* within the project *Topophonie*, ANR-09-CORD-022, http://topophonie.fr.

7. REFERENCES

[1] M. Casey, "Acoustic lexemes for organizing internet audio," *Contemporary Music Review*, vol. 24, no. 6, pp. 489–508, Dec. 2005.

[2] D. Schwarz, "Corpus-based concatenative synthesis,"

IEEE Signal Processing Magazine, vol. 24, no. 2, pp. 92–104, 2007.

[3] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[4] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE Trans. Image Processing*, vol. 11, pp. 146–158, 2002.

[5] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895 – 2907, 2003.

[6] G. Wichern, J. Xue, H. Thornburg, and A. Spanias, "Distortion-aware query-by-example for environmental sounds," in *Proc. WASPAA*, 2007, pp. 335–338.

[7] J. Xue, G. Wichern, H. Thornburg, and A. Spanias, "Fast query by example of environmental sounds via robust and efficient cluster-based indexing," in *Proc. ICASSP*, vol. 49, no. 3, 2008, p. 58.

[8] N. Misdariis, A. Minard, P. Susini, G. Lemaitre, S. McAdams, and E. Parizet, "Environmental sound perception: Metadescription and modeling based on independent primary studies," *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[9] V. Peltonen and V. Peltonen, "Computational auditory scene recognition," in *Proc. ICASSP*, 2001, pp. 1941–1944.

[10] D. Schwarz, "State of the art in sound texture synthesis," in *Proc. of Dafx11*, 2011, pp. 221–231.

[11] J. H. McDermott, A. J. Oxenham, and E. P. Simoncelli, "Sound texture synthesis via filter statistics," in *Proc. WASPAA*, 2009, pp. 297–300.

[12] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Synthesizing sound textures through wavelet tree learning," *IEEE Comput. Graph. Appl.*, vol. 22, no. 4, pp. 38–48, Jul. 2002.

[13] D. O'Regan and A. Kokaram, "Multi-resolution sound texture synthesis using the dual-tree complex wavelet transform," in *Proc. EUSIPCO*, 2007.

[14] G. V. D. Wouwer, P. Scheunders, and D. V. Dyck, "Statistical texture characterization from discrete wavelet representations," *IEEE Trans. Image Process.*, vol. 8, pp. 592–598, 1999.

[15] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 674–693, 1989.

[16] T. J. Ypma, "Historical development of the newton-raphson method," *SIAM Review*, vol. 37, no. 4, pp. 531–551, 1995.

[17] T. M. Cover and J. A. Thomas, *Elements of information theory*, New York, NY, USA, 1991.

[18] R. Kwitt and A. Uhl, "Lightweight probabilistic texture retrieval," *Trans. Img. Proc.*, vol. 19, no. 1, pp. 241–253, Jan. 2010.

[19] L. Bombrun, Y. Berthoumieu, N.-E. Lasmar, and V. Geert, "Multiscale colour texture retrieval using the geodesic distance between multivariate generalized gaussian models," in *Proc. ICIP*, 2008.

[20] M. Slaney, "Auditory toolbox," 1994.