REPRESENTING ENVIRONMENTAL SOUNDS USING THE SEPARABLE SCATTERING TRANSFORM

Carlo Baugé[⋆] *Mathieu Lagrange*[⋆] *Joakim Andén*[†] *Stéphane Mallat*[⋆⋆]

carlo.bauge@ircam.fr
[⋆]IRCAM (STMS CNRS UPMC) 1, place Igor Stravinsky 75004 Paris, France
[†]CMAP, Ecole Polytechnique Route de Saclay, 91128 Palaiseau, France
[⋆⋆]Ecole Normale Supérieure 45, rue d'Ulm 75005 Paris, France

ABSTRACT

Environmental sounds are an interesting subject of study for machine audition because of their wide variety of acoustical characteristics and their central presence in our everyday life. They are perceived effortlessly in the human auditory system whereas state-of-the-art computational systems are far from reaching the same efficiency.

In this paper we propose a novel representation of such sounds based on the scattering transform which has the property of stability to time-warping deformations and invariance to time-shift useful for classifications tasks. This representation is compared to several state-of-the-art approaches for the task of quantifying similarity between environmental sounds.

***Index Terms*—** scattering, separable scattering, wavelet, environmental sounds, similarity

1. INTRODUCTION

Environmental sounds have been less studied than other types of sounds like music [1] or speech [2]. Even though much is left to discover in the human auditory system, several modes of listening seem to be present, each one being presumably tuned to different types of input.

For example, in [3] Gaver distinguishes between two modes of listening used in different contexts and degrees of attention. The first one, called ”everyday listening”, corresponds to the way we usually hear in the day-to-day. Importance is given to source identification, localization in order to recognize which sources to interact with and which to avoid. This is opposed to the ”musical listening” mode, which emphasizes the detailed characteristics of the sound itself like timbre or pitch.

The latter is therefore more analytical and precise than the former, which is rather used to compute a fast signature of what we hear. For environmental sounds, everyday listening seems the most appropriate type of listening. How can we design a representation that would be useful for implementing this mode of listening? What kind of acoustic features should be kept in the representation and which should be discarded?

To provide some answers to those questions, we propose a new computational approach based on the scattering transform [4] that has the property of time-shift invariance. In Section 2, we review the state-of-the-art in terms of sound representations and similarity computation. Section 3 presents the scattering transform and introduces a variant that adds frequency transposition property. Sections 4 and 5 describes the experimental protocol used to compare the representations and discusses the results.

2. RELATED WORK

The spectrogram, that is the modulus of the Short-Term Fourier Transform (STFT), is a commonly used representation of audio signals as it decomposes the signal into a time/frequency plane which has local time-shift invariance and loses little information when discarding the phase information. In order to better match this representation with the output of the cochlea and gain stability with respect to deformation, the frequency axis is often scaled logarithmically, for example using the mel scale.

In most audio processing applications, the features are fed to a classifier that operates better on low-dimensional and decorrelated features. It is therefore convenient to apply a final processing step to transform the logarithm of mel-scale spectra. Even though the Principal Component Analysis (PCA) could be used for this stage, similar decorrelation properties can be obtained (at least on speech and music data) by projecting the log-mel spectra on a cosine basis, which leads to the well-known Mel-Frequency Cepstral Coefficients (MFCC) [2, 5].

Whatever the features chosen, their values evolve in time and this evolution carries information important to classifica-

This work has been partially supported by ANR-11-JS03- 005-01and ANR grant BLAN 012601. Part of this work has been done while M. Lagrange was visiting IRCCYN (UMR 6597).

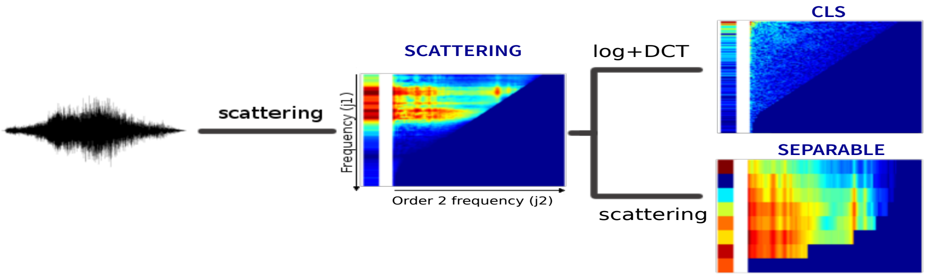

Fig. 1. Overview of the computation of CLS and separable scattering representations

tion. Several approaches exist in order to compare the representations of two sounds. In order to account for possible time-stretching between sounds, one can consider the Dynamic Time Warping (DTW) approach [2]. It aligns the two representations by finding an optimal match (with respect to cosine distance) between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" non-linearly in the time dimension to calculate a similarity measure independent of non-linear variations in the time dimension. Unfortunately, this non-linear matching has a complexity which is quadratic in the order of the number of frames in an audio sample.

A computationally less expensive approach is to model the diversity of the observations without preserving any sequentiality information. The observations are then considered as a Bag Of Frames (BOF) [6] which is approximated using mixture models representing the probability distribution of the observations in the feature space. For that purpose, the Gaussian Mixture Model (GMM) is a convenient choice and has been considered for a wide variety of tasks [7].

3. PROPOSED METHOD

Ideally, one would have a computational paradigm where the timing relationship between events is explicitly modeled at the feature level which would lead to simple and efficient processing at the decision level, be it for similarity computation or for classification. The scattering transform[1] [4], [8] yields such representation by cascading a wavelet filter bank and a modulus rectifier and has time-shift invariance property and stability to time-warping deformations.

3.1. Scattering operator

We consider a wavelet filter bank $(\psi_j)_{j<J+P}$ constructed by dilating a wavelet ψ whose octave bandwidth is $1/Q$.

$$\psi_j(t) = a^{-j}\psi(a^{-j}t) \text{ with } a = 2^{1/Q} \text{ and } j < J \quad (1)$$

The filter bank consists of J logarithmically spaced filters where each ψ_j covers the frequency interval $[2Q\pi a^{-j} - \pi a^{-j}, 2Q\pi a^{-j} + \pi a^{-j}]$. In the frequencies below $2Q\pi a^{-J}$, the filter bank consists of P (where $P \simeq 1/\ln(a)$) linearly spaced filters (for $J \leq j < J+P$) having the same frequency bandwidth as ψ_J, which is $2\pi a^{-J}$.

To compute the first-order scattering coefficients, we apply the filter bank followed by a modulus rectifier to a signal x yielding a scalogram. We obtain the first-order coefficients by applying a low-pass filter Φ_J to each frequency bin of the scalogram. This process is expressed in (2). Low-pass filter Φ_J accounts for frequencies below the lowest frequency of the filter bank. We obtain a representation similar to a mel-spectrum.

$$|x \star \psi_j| \star \Phi_J(t),\ \forall j < J+P \quad (2)$$

To recover lost information we compute what is called second-order coefficients. To do so, we consider each output $|x \star \psi_{j_1}(t)|$ for all $j_1 < J+P$ as a new signal and apply the same process as for the first order:

$$||x \star \psi_{j_1}| \star \psi_{j_2}| \star \Phi_J(t),\ \forall j_1, j_2 < J+P \quad (3)$$

The second-order coefficients (3) give a frequential representation of the amplitude modulations over time for each frequency bin defined by the filter bank. For a signal of length N, there are $Q\log_2(N/Q)$ first-order coefficients and $Q^2/2\log_2^2(N/Q^2)$ second-order coefficients.

Scattering has two relevant properties for us. The first is having the local time-shift invariance while keeping the amplitude modulations. Indeed, the information of the amplitude modulations inside a large time-window is lost in first-order coefficients but captured by second-order coefficients. The second property is the stability of the representation when the signal is slightly stretched in time as the logarithmic filter bank allows us to stabilize the high frequencies [4]. This approach share some of those properties with the spectrotemporal analysis proposed by Shamma &al [9].

The scattering representation is the basis for the representations we considered for our experiments, which apply a transform on the scattering coefficients to add frequency transposition invariance.

[1] MATLAB software for scattering and CLS can be found on `http://www.cmap.polytechnique.fr/scattering`

3.2. Cosine Log Scattering (CLS)

Computing Cosine Log-Scattering [8] from scattering is similar to calculating MFCC from the mel spectrum. MFCCs are widely used on musical and voiced sounds, as they enable to easily separate the two components in a source-filter model: $x(t) = e \star h(t)$. As mentioned earlier, the Discrete Cosine Transform (DCT) tends to concentrate the variance of the coefficients towards the low DCT frequencies.

Similarly to MFCCs, source and filter components can be linearly separated in the CLS representation. First-order CLS coefficients are computed by applying a DCT on the logarithm of first-order scattering coefficients. To compute second-order CLS coefficients at a time t_1, we consider the logarithm of second-order scattering coefficients: $\log(||x \star \psi_{j_1}| \star \psi_{j_2}| \star \Phi_J(t_1))$. A DCT is applied first along j_2 and then along j_1 yielding a representation parametrised by k_1 and k_2. These DCTs efficiently decorrelate the scattering coefficients [8].

By discarding high DCT-frequency coefficients, invariance to transposition and more stability to deformations is achieved. However, there is an intrinsic trade-off between loss of information and stability using this approach.

3.3. Separable scattering

To improve this trade-off, we propose to replace the log-DCT step by the application of the scattering transform anew on the scattering representation. Loss of information then does not come from discarding coefficients but from averaging them. We term this new form of scattering "separable" as it imply two separate steps of scattering. A first property is that this representation is invariant to frequency transposition within the support of the lowpass filter in the second scattering transform.

In addition, as the logarithmic scale in the scattering filter bank ensured stability to time deformations in the scattering representation, this same property is obtained for frequency deformations when the first-order scattering transform is applied anew to the scattering representation itself.

More precisely, a scattering transform is applied as would a log-DCT on first-order scattering coefficients. On second-order coefficients the scattering transform is only applied along j_1, in the direction of the dimension of the acoustic frequency which lead to a filtering operation similar to the one considered in [10].

4. EXPERIMENTS

4.1. Databases

We considered 3 databases [11], [12] consisting of short environmental sounds (up to 10 seconds) all sampled at 44.1 kHz. The first (called "Gygi") contains 100 sounds organised in 50 classes, each of 2 sounds. The sounds cover a wide variety of environmental sounds, ranging from baby cries to airplane noises to footsteps to harp sounds. Within this database, the number of sounds per class was found to be too low compared to the diversity of the classes. We therefore decided to widen it to obtain a minimum of 4 sounds per class which lead to the "Gygi extended" database containing 294 sounds organised in the same 50 classes. The last database is called Houix (from [12]) and contains more specific sounds that occur in a kitchen. The 60 sounds are organised here into 4 classes (solid, gas, liquid and machine).

	Durations(s)		**Elements/cl**		**Total**	
	$mean$	σ	$mean$	σ	Cl.	Elts.
$Gygi$	2.32	0.86	2.00	0.00	50	100
$GygiExt$	2.40	0.92	5.88	1.77	50	294
$Houix$	2.92	1.98	15.00	13.83	4	60

Table 1. Statistics of the 3 sound databases. These refer to the sound durations and the sound classes. Mean and standard deviation are shown. The last two columns show the total number of classes and elements.

Table 1 sums up the characteristics of the 3 databases.

4.2. Methods

For all the methods considered in this study, the amplitude of the sounds is first normalized using the Root Mean Square (RMS) in a 100-millisecond window around the peak amplitude of each sound.

The state-of-the-art representations used are the spectrogram, the mel-spectrum and the MFCCs. To compute the distances between sounds the Bag of Frames (BOF) algorithm is used [6]. It learns a GMM model [7] for each sound and approximates the Kullback-Leibler divergence to compute a distance between sounds [2]. A second distance algorithm used is the DTW which takes into account more temporal information on the sounds[3].

For comparison, the scattering transform is applied on each sound with J such that Φ_J will completely average the sound. This, in order to guarantee time-translation invariance of the representation. To obtain the same dimensionality for each sound, all sounds are zero-padded. This gives a vector length of 262144, 262144, and 524288 samples for the Gygi, GygiExt and Houix databases. In our case we used a quality factor Q of 8 in the scattering filter bank.

The CLS representation is computed by retaining only the first 10% lowest DCT-frequency first-order coefficients 10% of the second-order coefficients along k_1 and the 3 lowest DCT-frequency coefficients along k_2. Indeed, performances are found to be maximal around this ratio.

[2] MATLAB software for BOF can be found on `http://www.jj-aucouturier.info/projects/mir/boflib.zip`

[3] MATLAB software for DTW can be found on `http://labrosa.ee.columbia.edu/matlab/dtw/`

	$RAND$	BOF	DTW	CLS_1			CLS_{1+2}			Sep_1			Sep_{1+2}		
				L^1	L^2	cos.	L^1	L^2	cos.	L^1	L^2	cos.	L^1	L^2	cos.
$Gygi$	5.1	31.8	25.8	26.8	23.3	21.6	37.2	31.9	33.1	19.9	20.8	30.0	35.8	35.4	**44.4**
$GygiExt$	3.6	27.9	19.3	19.4	17.6	19.4	28.4	23.0	25.4	20.9	20.7	20.9	35.6	32.2	**38.9**
$Houix$	43.6	54.6	55.5	55.0	54.8	55.0	53.2	52.0	52.6	57.4	57.3	52.0	56.1	56.8	**59.0**

Table 2. Mean average precision results (in percentage) on each database for different algorithms. From left to right: random distances, state-of-the-art algorithms BOF and DTW, order 1 CLS and order 1+2 CLS, order 1 separable scattering and order 1+2 separable scattering.

The separable scattering representation is obtained by applying the scattering transform with a quality factor of 1 and again averaging over the whole frequency-signal. This means that the representation is completely invariant to frequency transposition.

The global representation has the form of a vector containing both first-order and second-order coefficients. Three standard distance measures are considered: L^2, L^1 norm, and cosine distance. Performance is evaluated by computing the Mean Average Precision over the resulting distance matrix [13].

5. RESULTS

Results are shown in Table 2. The first two columns show the results for a random distance matrix and the best results for the two state-of-the-art algorithms. To compute state-of-the-art algorithms we varied sound representation (spectrogram, mel spectrum and MFCCs) and for BOF algorithm the number of Gaussians used per model varied from 1 to 20. Due to the short duration of the audio samples, the models that gave the best results were those with a low number of Gaussians, frim 1 to 3 depending on the database. The four last columns show the results for CLS and separable scattering when considering only first-order coefficients (CLS_1 and Sep_1) or both first- and second-order coefficients (CLS_{1+2} and Sep_{1+2}).

If we compare the results of first-order with those of order $1 + 2$, results are systematically better with order 1+2 for both CLS and separable representations. This indicates the importance of taking into account the amplitude modulations in the representation of sounds. As far as the kind of distance metric is concerned, the L^1 norm compares favorably to the others for CLS and the cosine distance for separable scattering. When considering first- and second-order separable scattering coefficients, cosine distance performs systematically better than the other distances.

The separable scattering also consistently shows better results than the CLS representation. This leads us to believe that the properties of invariance and stability of the separable scattering are useful in this task of similarity between environmental sounds.

More generally, the separable scattering representation shows better results than the two state-of-the-art algorithms with a stronger difference (+12.6% and +11%) for the two databases containing a wider variety of sounds (Gygi and GygiExt).

To summarize, scattering representations and in particular separable scattering seem to be more appropriate for our task than the usual representations such as mel-spectrum or MFCCs. It shall be noted that the distance measures we used for the two scattering representations were conceptually and algorithmically simpler than those of BOF and DTW.

6. CONCLUSION

We introduced a novel representation of environmental sounds by adding a property of invariance to frequency transposition in the scattering representation. Experimental results demonstrate its validity in the task of measuring similarity between environmental sounds on 3 databases.

Future works will consider other types of sounds such as speech and music in order to better understand what kind of invariance is needed along the time and frequency axes with respect to the specificity of the data and the task at hand.

7. REFERENCES

[1] G. Tzanetakis and P. R. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs: Prentice Hall, 1978.

[3] W. W. Gaver, "What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, Mar. 1993.

[4] S. Mallat, "Group invariant scattering," *Communications in Pure and Applied Mathematics*, to appear, http://arxiv.org/abs/1101.2286.

[5] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *In International Symposium on Music Information Retrieval*, 2000.

[6] J.-J Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for poly-

phonic music," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[7] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*, pp. 659–663. 2009.

[8] J. Andén and S. Mallat, "Multiscale scattering for audio classification," in *ISMIR*, 2011, pp. 657–662.

[9] Taishih Chi, Powen Ru, and Shihab a. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887, 2005.

[10] M Deviren and K Daoudi, "Frequency and wavelet filtering for robust speech recognition," *Artificial Neural Networks and Neural Information Processing*, p. 452460, 2003.

[11] B. Gygi, Gary R. Kidd, and CS Watson, "Similarity and categorization of environmental sounds," *Perception And Psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

[12] O. Houix, G. Lemaitre, N. Misdariis, P. Susini, and I. Urdapilleta, "A lexical analysis of environmental sound categories," *Journal of Experimental Psychology: Applied*, 2012.

[13] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.