Auditory Sketches: Sparse Representations of Sounds Based on Perceptual Models

Clara Suied[1,3,⋆], Angélique Drémeau[2,4,⋆], Daniel Pressnitzer[1], and Laurent Daudet[2]

[1] Laboratoire Psychologie de la Perception, CNRS - Université Paris Descartes & Ecole Normale Supérieure, 29 rue d'Ulm, 75230 Paris Cedex 5, France
clara.suied@irba.fr, Daniel.Pressnitzer@ens.fr
[2] Institut Langevin, ESPCI ParisTech and CNRS UMR 7587, Université Paris Diderot, 1 rue Jussieu, 75005 Paris France
angelique.dremeau@telecom-paristech.fr, laurent.daudet@espci.fr
[3] Institut de Recherche Biomédicale des Armées (IRBA), Département Action et Cognition en Situation Opérationnelle, 91223 Brétigny sur Orge, France
[4] Institut Mines-Telecom - Telecom ParisTech - CNRS/LTCI UMR 5141, 75014 Paris

Abstract. An important question for both signal processing and auditory science is to understand which features of a sound carry the most important information for the listener. Here we approach the issue by introducing the idea of "auditory sketches": sparse representations of sounds, severely impoverished compared to the original, which nevertheless afford good performance on a given perceptual task. Starting from biologically-grounded representations (auditory models), a sketch is obtained by reconstructing a highly under-sampled selection of elementary atoms. Then, the sketch is evaluated with a psychophysical experiment involving human listeners. The process can be repeated iteratively. As a proof of concept, we present data for an emotion recognition task with short non-verbal sounds. We investigate 1/ the type of auditory representation that can be used for sketches 2/ the selection procedure to sparsify such representations 3/ the smallest number of atoms that can be kept 4/ the robustness to noise. Results indicate that it is possible to produce recognizable sketches with a very small number of atoms per second. Furthermore, at least in our experimental setup, a simple and fast under-sampling method based on selecting local maxima of the representation seems to perform as well or better than a more traditional algorithm aimed at minimizing the reconstruction error. Thus, auditory sketches may be a useful tool for choosing sparse dictionaries, and also for identifying the minimal set of features required in a specific perceptual task.

1 Introduction

Sound signals are one-dimensional time series, reflecting the variation of acoustic pressure in the air. There is a variety of ways to represent such time-series,

⋆ These authors contributed equally to this work.

M. Aramaki et al. (Eds.): CMMR 2012, LNCS 7900, pp. 154–170, 2013.

starting with Fourier transforms or wavelet analyses [1]. Each representation is defined in a set of basis functions on which the time-series are projected: complex exponentials for the Fourier analysis, or dilated and translated versions of a mother wavelet for wavelets. In an "atomistic" view of this analysis process [2], the set of basis functions is often called the "dictionary", and its elements the "atoms". Desirable properties for a dictionary may be the orthogonality between elements, or its completeness and invertibility (*i.e.,* it is possible to represent any signal and transform it back without any loss of information). More recently, for applications such as source separation or denoising, further properties have been shown to be useful, such as sparsity (see [3] for a review), where only a few non-zeros coefficients can be used to represent a signal. In practice, exact sparsity is never achieved for sound signals, but still most of them can be well approximated by sparse representations (the approximation error decays quickly as the number of terms increases), a property often referred to as *compressibility*. Such sparse representations are usually computed through some non-linear algorithms, optimizing a balance between sparsity and data fidelity [4].

The size and nature of the (possibly over-complete) dictionary must be carefully chosen, as larger dictionaries tend to provide sparser representations, but the computational cost of the associated estimation algorithms may become prohibitive, and high coherence in the dictionary elements may result in identifiability issues. The choice of the dictionary elements, or "atoms", is also of prime importance, as these must be designed to fit local features of the signals under study ; they can be chosen *a priori* or learnt on the data itself [5].

In this paper, we outline an original method for investigating sparse representations of sound signals, based on perceptual considerations. The underlying idea is simple: sounds are not just any time-series, they are time-series that are being perceived by listeners. As a consequence, not all information in sound is relevant for a given listening task. For instance, speech content is remarkably resilient to large acoustic distortions [6], showing that a massive information-loss can be tolerated for tasks like speech intelligibility in quiet. The key is that the distortion should preserve a small but sufficient set of features for the task. Here we introduce the metaphor of an "auditory sketch": a sketch is a signal that has been severely impoverished compared to the original sound, and thus is clearly distinguishable from it, but that still retains enough of the original critical features to afford good performance on a target task.

A schematic of the work flow we suggest to obtain auditory sketches is presented on Fig. 1. The method is iterative, and places the listener at the centre of the design loop. The first proposal is to use auditory models. Auditory models refer to a class of signal-processing algorithms trying to mimick the way the acoustic signal is transformed along the human auditory pathways. For instance, the cochlea performs a time-frequency decomposition, which can be approximated to a first degree by a bank of overlapping band-pass filters [7]. The resulting representation is often termed an "auditory spectrogram". Subsequent stages of processing in the auditory pathways display more complex processes, which are

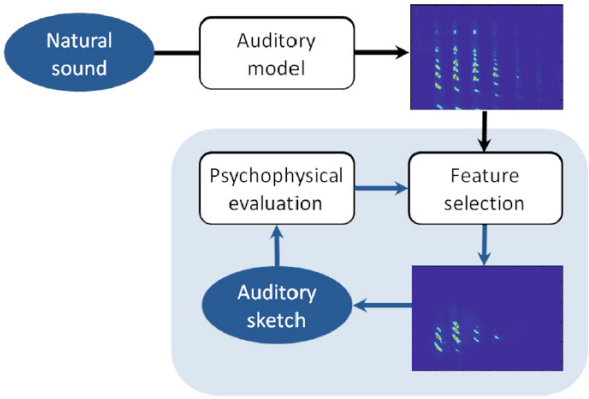

Fig. 1. Overview of the sketch design method. An auditory representation of a natural sound is generated (in this example, an auditory spectrogram) and only a few features are retained. The auditory model is then inverted for re-synthesis of the candidate sketch. Psychophysical experiments involving human listeners are then used to evaluate the efficiency of the selected features. The process is repeated iteratively to discover a sparse set of features that afford good performance with sound class and task at hand.

currently only poorly understood. For instance, neurons in the primary auditory cortex exhibit a variety of selectivity to spectro-temporal features such as spectral, temporal, or joint-spectral temporal modulations. Models nevertheless exist to idealise such a processing as a bank of 2-D wavelets operating on the auditory spectrogram [8]. Such schematic "cortical" representations have been shown, for instance, to be sufficiently rich to be an efficient front-end for timbre classification [9].

It is hoped that, because they are inspired by the physiology of the human ear, such auditory representations will contain the features that are relevant to perception. However, these representations are massively over-complete, so it is not obvious to assess which part of the representation is relevant for a given task. This is where we use a second step in the sketches method: the representations are sparsified by keeping only a small set of non-zero coefficients. A variety of selection algorithms can be envisioned, as discussed below.

Finally, to check that the relevant features have been preserved, we invert the sparse representations back into sound signals. The resulting sounds are then used in psychophysical tests with human listeners. The process should be repeated iteratively until the selection of sparse features affords good performance on the target perceptual task.

In this paper, we present preliminary data as a proof of concept for the sketches process. We compare two different auditory models, aimed at representing two distinct stages of auditory processing: the auditory spectrogram and the cortical representation [8]. The selection of non-zero coefficients from the models is obviously a central issue, and here we compare two potential candidates: a simple peak-picking algorithm, and an analysis-based iterative thresholding method

[10]. Finally, the psychophysical task chosen is that of recognition of emotion in short sound snippets. Sounds are extracted from a calibrated database of natural emotional signals [11], transformed as sketches, and then listeners have to identify the original emotion in a forced-choice task (happiness, anger, sadness, disgust). Only the first iteration in the method is tested.

2 Sparse Representations of Sounds: Dictionaries and Algorithms

The "sketching" problem we are interested in can be formalized as follows. We look for the sketch $\mathbf{x} \in \mathbb{R}^N$, representation of the audio signal $\mathbf{y}$ such as

$$\mathbf{y} = \mathbf{x} + \epsilon, \tag{1}$$

where ϵ stands for the difference between the original audio signal $\mathbf{y}$ and its sketch $\mathbf{x}$. Within our study, the sketch $\mathbf{x}$ is then assumed to have a sparse representation in a given dictionary.

Traditionally, the quality of the sparse representation is measured both in terms of sparsity and approximation (*i.e.,* the fidelity to the original signal). It depends on the dictionary in which the decomposition is performed, and the procedure for the selection of sparse features (and the corresponding algorithms). Here, an additional stage is considered. Following the algorithmic procedure implementing the sparse decomposition, the appropriateness of the resulting sketch to the target task is further tested through a psychophysical evaluation (see Fig. 1). Ideally, the whole procedure is then iterated to refine both the dictionary and the procedure for the selection of sparse features (in terms of objective functions, sparsity levels and algorithms). In this section, we discuss *a priori* choices for the dictionary, in Subsect. 2.1, and the decomposition procedures, in Subsect. 2.2. These can be thought of as reasonable initial conditions for the sketches process. In the context of this paper, they also serve to illustrate the potential of the method.

2.1 Auditory-Motivated Dictionaries

The choice of the dictionary is deeply related to the targeted application. In denoising tasks, for example, emphasis may be put on the match to the characteristics of the signal itself. Here, we will favour biologically-inspired dictionaries that take into account the ear physiology. The underlying hypothesis is that perception is shaped by the neural processing of sound. For instance, the frequency selectivity observed in auditory masking (which part of the sound will effectively be detected by a listener) is thought to be linked to frequency selectivity on the cochlea.

We chose to use the auditory model described by Chi *et al.* [8] and freely available as the "NSL toolbox"[1]. As mentioned in the introduction, the model

[1] `http://www.isr.umd.edu/Labs/NSL/Software.htm`

includes both an auditory spectrogram and a "cortical" spectro-temporal analysis of the spectrogram. It has proved successful for several signal-processing applications, such speech intelligibility assessment [12], or computational modeling of timbre perception [9].

The model consists of two major auditory transformations:

i) The *early stage* transforms the one-dimensional acoustic waveform to a two-dimensional pattern obtained with a bank of constant-Q filters, followed by spectral sharpening (lateral inhibition) and compression. Fig. 2 illustrates the result of such a transformation, producing what is termed an *auditory spectrogram.*

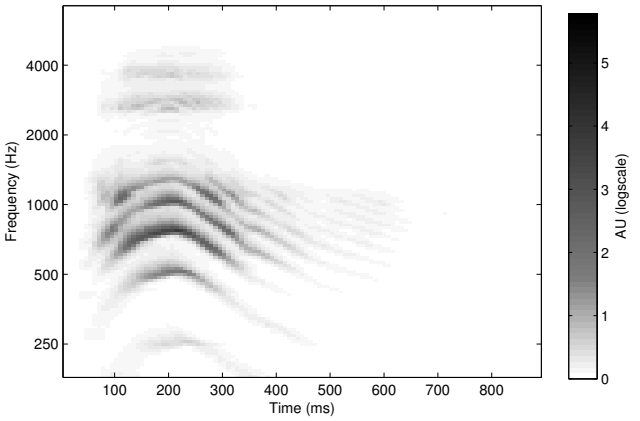

Fig. 2. Example of an auditory spectrogram (AU: arbitrary units, log scale). The sound analyzed is a short affect burst expressing anger [11]. The voiced quality of the sound is visible in the harmonic structure of the frequency components, which are themselves shaped by the vocal formants. A continuous glide of the fundamental frequency (up then down) is also salient.

ii) The *cortical stage* implements then a more complex spectrotemporal analysis, presumed to take place in the mammalian primary auditory cortex. The transformation relies on a bank of filters, selective to different spectrotemporal modulation parameters which range from slow to fast rates temporally and from narrow to broad scales spectrally. It results in a four-dimensional, time-frequency-scale-rate representation, referred to as the *cortical representation* of the signal. A detailed description of such a representation is beyond the scope of the paper, the reader is refered to [8]. Fig. 3 nevertheless illustrates some features of the cortical representation.

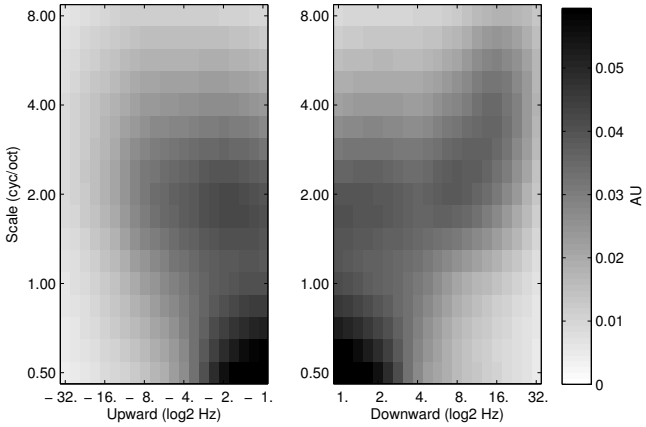

Fig. 3. Example of a cortical representation (AU: arbitrary units). The sound is the same as in Fig. 2. We only illustrate the projection of the 4-D cortical representation on the "rate" and "scale" dimensions (the cortical representation was averaged over time and over frequency channels). The pattern of rate and scale coefficients describe the spectro-temporal evolution of the sound. For instance, because the fundamental frequency glide induces temporal amplitude modulations in many frequency channels, there is a range of non-zero modulation rates in the representation. The left and right panels are for upward and downward spectro-temporal modulations, respectively (see [8] for details).

Because our method relies on a listening test, an important issue is the invertibility of the representations used. If phases are preserved, the (standard or auditory) spectrograms are easily invertible, akin to the overlap-add resynthesis procedure of the standard spectrogram. However, if non-linear processing makes phase information meaningless, as is the case here (lateral inhibition, thresholding, compression), perfect reconstruction cannot be achieved.

In order to obtain time-domain signals that are compatible with the spectrogram, one can resort to phase estimation algorithms that exploit the intrinsic redundancy of the transforms, such as the Griffin and Lim [13] phase reconstruction iterative procedure, or improvements thereof (see [14] for a review). It should be noted that this algorithm reconstructs a set of phases that are consistent, but that may be completely different from the original phases, thus precluding any time-domain sample-by-sample comparison. Here, we use the method of [15], developed for auditory spectrograms and which provides reconstructions that are highly perceptually similar to the original signal, whenever the auditory spectrogram is not modified.

The parameters chosen for the model of [8] were as follows. The audio signals were sampled at 16kHz. The auditory spectrogram was obtained with a bank of 128 bandpass filters and 8-ms time windows. The cortical stage had 5 rate channels for temporal modulations (from 1 to 32Hz) and 6 scale channels for spectral modulation (from 0.5 to 8 cycles/octave), resulting in a redundant representation 60-times larger than the original signal.

2.2 Sparsification of Auditory Models

The next step in the design of sketches is the choice of a selection procedure for the features. Here again many choices are possible. Note that the iterative method of Fig. 1 is conceived precisely as a way to refine the selection process. As a first step, to gain some insight into the kind of methods that could serve as initial choices in the iterative process, we compare two selection procedures contrasting two different approaches:

- Algorithm IHT (iterative hard thresholding), based on a sparse analysis scheme
- Algorithm PP (peak-picking), based on peak-picking of local maxima

It is important to stress that, as we shall discuss, these two procedures are not just different from an algorithmic point of view. More importantly, one of them aims at optimizing the quadratic reconstruction error (IHT), while the other (PP) is purely feedforward and does not include any optimization step. In both cases, the ultimate success of the selection or otherwise is estimated by means of the perceptual task.

Algorithm IHT: Sparse Analysis by Iterative Hard Thresholding. Two mathematical sparsity formalisms are possible, according to the adopted – *analysis* or *synthesis* – approach. On the one hand, from the analysis point of view and within our sketching problem, the sketch $\mathbf{x}$ is assumed to produce a sparse output, which can be expressed under a matrix formulation as

$$\mathbf{z} = \mathbf{A}\mathbf{x}, \tag{2}$$

where $\mathbf{z} \in \mathbb{R}^M$ is sparse, *i.e.,* contains few non-zero elements, and $\mathbf{A}$ is a $(M \times N)$-matrix with $M \geq N$ representing the analysis operator. On the other hand, from the synthesis point of view, the sketch $\mathbf{x}$ is seen as the sparse combination of atoms, namely

$$\mathbf{x} = \mathbf{D}\mathbf{z}, \tag{3}$$

where $\mathbf{D}$ is a $(N \times M)$-matrix with $M \geq N$ representing the dictionary, and $\mathbf{z}$ is sparse.

Within the sparse-representation framework, the synthesis approach constitutes the most common formalism, being the subject of numerous contributions (see *e.g.*, [16] for a review of the algorithms dealing with synthesis sparsity). However, as described above, the representations we chose rely on a sequence of filters applied to the signal and analyzing their outputs, which tends to favor the analysis point of view.

Furthermore, the sparsity constraint in which we are interested in is not taken into account in the same way within both formalisms. The synthesis formulation, by its generative nature, leads potentially to a greater compactness of the signal. But, with this formulation, the choice of the atoms to represent the signal has

huge implications: a wrong decision may cause the selection of additional wrong atoms as compensation. This is not the case with the analysis formulation, where all atoms contribute equally to the representation of the signal [17]. We will thus adopt the analysis point of view in the remain of the paper. Hence, depending on the processing level, a sketch $\mathbf{x}$ of the audio signal $\mathbf{y}$ is built from a sparse auditory spectrogram or a sparse cortical representation of the signal $\mathbf{y}$.

Considering the analysis formulation (2), the estimation of the sketch $\mathbf{x}$ can then be expressed as

$$\mathbf{x}^\star = \underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{x}||_2^2 \text{ subject to } ||\mathbf{A}\mathbf{x}||_0 \leq L, \tag{4}$$

where $\|.\|_0$ denotes the ℓ_0 pseudo-norm, counting the number of non-zero elements, and L is a parameter specifying the maximum number of non-zero elements in $\mathbf{z}$.

Finding the exact solution of (4) is an NP-hard problem, *i.e.,* it generally requires a combinatorial search over the entire solution space. Here, we use a suboptimal (but tractable) algorithm based on the iterative hard thresholding procedure introduced in [10]. This algorithm presents indeed several desirable properties:

i) Its implementation is very simple, in accordance with a filter-bank procedure, as considered within our model (see Subsect. 2.1).
ii) Its complexity is low, in $\mathcal{O}(N \log N)$, N being the number of iterations. This property is very valuable since the considered biologically inspired model involves complex mathematical computations, requiring thus a light integration procedure.

Note that the analysis-based IHT algorithm is different from the most standard synthesis-based iterative hard thresholding algorithms in the literature [18], often used in the framework of compressed sensing.

Algorithm PP: Peak-Picking of Local Maxima. The second algorithm considered in this paper is based on a simple local maxima detection.

The procedure, with variants already used in the literature (see *e.g.*, [19,20]), is based on a local gradient evaluation. In our case, the peak-picking was done on either the auditory spectrogram (finding 2-D local maxima) or the cortical representation (finding 4-D local maxima). The algorithm proceeds as follows : first, all local maxima (on the magnitude of the coefficients) are selected. Then, they are sorted by decreasing order and only the L largest are kept, L being related to the desired degree of sparsity. Note that this algorithm is not iterative, without any optimization procedure, and therefore is very fast.

It should also be noted that, as opposed to the vast majority of sparse decomposition/analysis algorithms, such as IHT described above, the goal of this analysis scheme is not to achieve the best approximation (in a least-squares sense) of the signal for a given number of coefficients. Instead, the rationale is that, if the representation itself is efficient, the selection mechanism can be

rather crude: within a zone of the parameters space, local maxima should express salient features.

3 Psychophysical Experiments

The core idea of the sketches process is to put the listener at the centre of the design procedure. Thus, as candidate sketches are obtained, they are used in a perceptual task where a performance measure can be obtained. If a high performance is observed, then this indicates that the set of features that have been selected in the sparsification process is sufficient for the task, even though the sketch itself may sound very different from the original signal.

We now report two experiments using a perceptual task of emotion recognition. We asked listeners to report whether a short vocal sound expressed happiness, sadness, anger, or disgust. Each emotion was represented by several sound samples, selected from a calibrated database [11]. The main aim was to provide a first test of the sketches approach: could listeners perform the task on sounds that were severely impoverished compared to the original? More precise questions as to the nature of the sketching process were asked in each experiment.

3.1 Experiment 1: Comparison of Two Auditory Representations

Rationale. Here we wanted to investigate the influence of the basic representation used to produce sketches. We used auditory models, but contrasted auditory spectrograms with spectro-temporal "cortical" representations. The robustness of sketches to the presence or absence of noise was also tested. Indeed, if we assume that the goal of the sketches is to identify perceptually-important features of sounds, a certain robustness to noise is desirable. Robustness to noise is thus one indication that the representation is well-suited to the sound class of interest. Finally, the sparsity that can be achieved with the method was evaluated: a better representation should produce a sparser code.

Material and Methods

Participants. There were 10 participants (6 men and 4 women), aged between 19 and 39 years ($M = 25.8$ years). All listeners had self-reported normal-hearing. They all provided informed consent to participate in the study, which was conducted in accordance with the guidelines of the declaration of Helsinki.

Stimuli. All sounds were derived from the Montreal Affective Voices database [11]. They consisted of recorded nonverbal emotional interjections (on the French vowel /*a*/). Among the available stimuli, we selected four emotions that were easily recognized (see [11]): anger, disgust, happiness, and sadness. Each emotional interjection was uttered by 10 different actors (5 male and 5 female). The original sounds had very different durations (from 0.4 s to 1.2 s), so we shortened some of the stimuli (happiness and sadness, mainly) to avoid recognition

cues linked to duration. The modified versions of the sounds were still easily recognized, as confirmed by an informal experiment. The modified sounds had an average duration of 0.99 s (std= 0.2). A repeated-measures ANOVA performed on the 40 sounds (4 emotions for the 10 speakers) revealed no significant differences between the mean duration of each emotion $[F(3, 27) = 0.95; p = 0.4]$. These 40 sounds constituted the baseline stimuli.

For the "noise" conditions, pink noise was added to the original sounds, with a signal-to-noise ratio of -6 dB.

The sketch process was performed either on the original sound or on the noise version of the sound. In this first experiment, it was only performed using the PP algorithm. Two auditory representations were compared: the auditory spectrogram and the cortical representation (see above). Three degrees of sparsity were also compared: 10, 100, and 1000 features/second were retained from the auditory representations. The measure of features/second, which we refer to as the degree of sketch, is only indirectly related to the quantity of information retained from the original signal (as for instance it ignores the size and nature of the dictionnary). However, it serves here as a first approximation of sparsity.

Apparatus. Stimuli were presented through an RME Fireface digital-to-analog converter at a 16-bit resolution and a 44.1 kHz sample-rate. They were presented to both ears simultaneously through Sennheiser HD 250 Linear II headphones. Presentation level was at 70 dB(A), as calibrated with a Bruel & Kjaer (2250) sound level meter and ear simulator (B&K 4153). Listeners were tested individually in a double-walled Industrial Acoustics (IAC) sound booth.

Procedure. A 4-AFC (Alternative Forced Choice) paradigm was used. In each trial, participants heard a single sound, which could be one of the 4 target emotions. They had to indicate whether the sound they just heard was a representative sound of happiness, sadness, anger, and disgust. Visual feedback was provided after each response.

14 conditions were presented in a randomized fashion to each participant, for a total of 1120 stimuli in total: original sounds vs. sketches and no noise vs. noise. For the sketches, we compared the auditory spectrogram vs. cortical representation and the degree of sketch (10, 100, or 1000 feat/s). The experiment lasted approximately 1 hour. The experiment was divided into small blocks, to allow time for breaks.

Results. Results are illustrated on Fig. 4. A first important observation is the overall good performance, well above the chance level (i.e. 25%), with a mean percent correct of 93% for the original sounds, and of 55% for the sketches sounds. A second key result rests upon the comparison of the two auditory models used to create the sketches: overall, the auditory spectrogram outperformed the cortical representation. Data were analyzed with a repeated-measures analysis of variance (ANOVA). We first evaluated the overall difference between the original sounds and the sketches, in the two noise conditions. A repeated-measure ANOVA revealed main significant effects for the type of sound (original

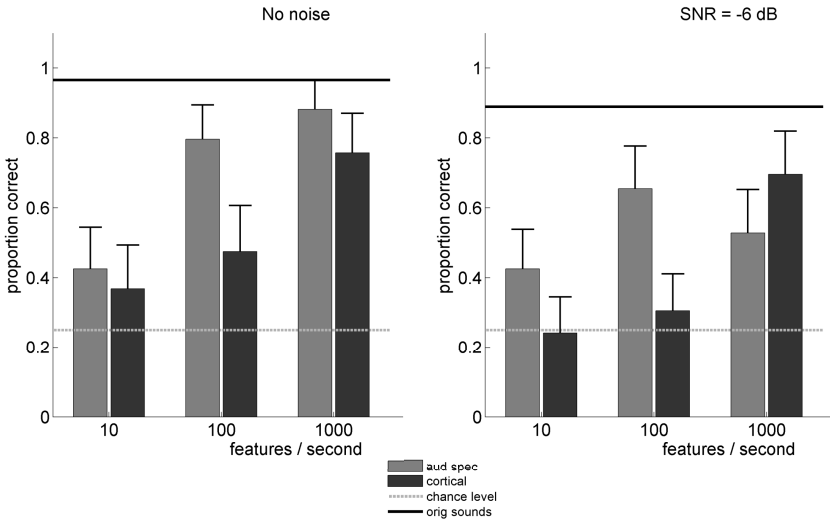

Fig. 4. Results for Experiment 1. Recognition performance of the sketches sounds corresponding to two different auditory models (aud spec for the auditory spectrogram, and cortical model), without (left panel) and with (right panel) noise. Error bars correspond to the standard error of the mean. Performance was overall higher for the auditory spectrogram than for the cortical model. These recognition data for the sketches sounds are compared to an upper baseline : the average recognition performance for the original sounds (black line). They are also compared to a lower baseline: the chance level, i.e. 25% here (dotted gray line).

vs. sketch) $[F(1,8) = 1172.55; p < 0.0001]$ and for the noise condition (silence vs. noise) $[F(1,8) = 441.81; p < 0.0001]$, as well as a significant interaction between these two variables $[F(1,8) = 21.66; p < 0.005]$. These results show that the overall recognition performance was better for the original sounds than for the sketches, and that, as expected, noise had a detrimental effect on performance; the influence of noise was more pronounced for the sketches than for the original sounds.

We then analyzed in more details data for the sketches sounds only. We performed a repeated-measure ANOVA with noise (silence vs. noise), model (auditory spectrogram vs. cortical), and features (10, 100, and 1000 feat/s) as within-subjects variables. It revealed main significant effects of noise $[F(1,8) = 582,23; p < 0.0001]$, model $[F(1,8) = 101,44; p < 0.0001]$, and features $[F(2,16) = 138,01; p < 0.0001]$. It also revealed significant interaction between features and model $[F(2,16) = 89,80; p < 0.0001]$, features and noise $[F(1,8) = 21,09; p < 0.0001]$, as well as a significant third-order interaction between features, model, and noise $[F(1,8) = 37,81; p < 0.0001]$. These results highlight that: performance was better in silence than in noise; performance increased as the number of features per second increased; the auditory spectrogram model led to better performance than the cortical model (with one notable

exception, that was responsible for the significant third-order interaction: in the noise condition, for 1000 feat/s, the cortical model led to better performances than the auditory spectrogram model).

3.2 Experiment 2: Comparison of Two Sparsification Algorithms

Rationale. Experiment 1 served as a first proof of concept of the sketches process: the overall recognition performance for sketches sounds was good (55%, i.e. well above the chance level). This was the case even though the selection algorithm, PP, was extremely crude and did not contain any optimization. Here, we compare the PP algorithm with a more traditional signal-processing approach, the IHT algorithm, that minimizes the reconstruction error (see Sect. 2.2).

Material and Methods

Participants. There were 10 participants (5 men and 5 women), aged between 19 and 34 years (M = 23.2 years). All listeners had self-reported normal-hearing. They all provided informed consent to participate in the study, which was conducted in accordance with the guidelines of the declaration of Helsinki.

Stimuli. Stimuli were very similar to Experiment 1, the only differences here being that: (i) only the auditory spectrogram was used as an auditory representation for the computation of the sketches; (ii) two sparsification algorithms were used to produce the sketches: IHT and PP (see Subsect. 2.2 for details).

Apparatus and Procedure. The apparatus was the same as in Experiment 1. The procedure was also very similar. Here, the 12 conditions that were presented in a randomized fashion to the participant were a combination of 3 parameters: type of algorithm (IHT vs. PP), noise (with or without), and degree of sketch (10, 100, and 1000 feature/second).

Results. Results of this second experiment are illustrated on Fig. 5. This second experiment confirms and reproduces some important results of Experiment 1: an overall good recognition performance, with a mean percent correct of 93% for the original sounds, and of 60% for the sketches sounds. It also shows that the PP algorithm generally outperformed the IHT algorithm. Similar analyses as for the Experiment 1 were conducted. Firstly, the overall ANOVA reproduced results of Experiment 1: performance was better for the original sounds than for the sketches $[F(1,9) = 708.77; p < 0.0001]$; performance was also better in silence that in the noise $[F(1,9) = 119.44; p < 0.0001]$. For this experiment as well, the detrimental effect of the noise was more pronounced for the sketches than for the original sounds [significant interaction between the type of sound and the noise condition: $[F(1,9) = 12:90; p < 0:006]$.
Secondly, a detailed repeated-measures ANOVA on the sketches only revealed that: as expected, performance was better in silence than in noise $[F(1,9) = 148.98; p < 0.0001]$; performance increased as the number of features per second increased $[F(2,18) = 283.89; p < 0.0001]$.

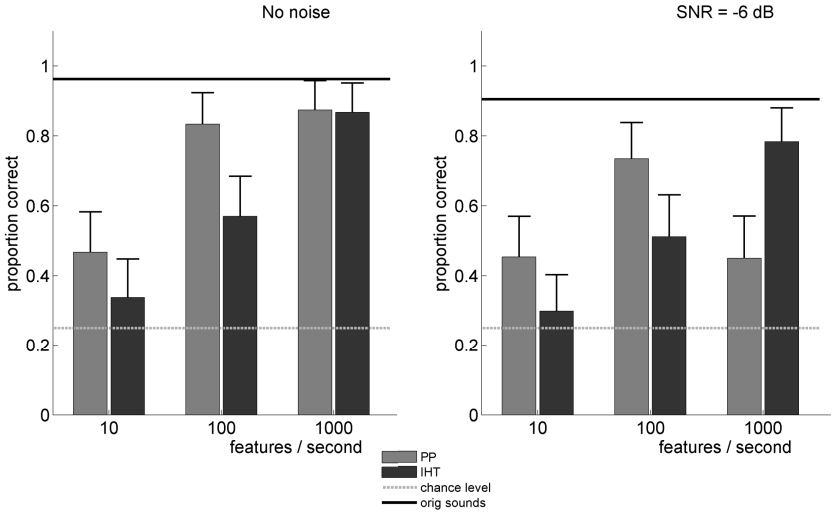

Fig. 5. Results for Experiment 2. Recognition performance of the sketches sounds corresponding to two different sparsifying algorithms (PP for peak-picking, and IHT for iterative hard thresholding), without (left panel) and with (right panel) noise. Error bars correspond to the standard error of the mean. Performance was overall higher for the PP than for the IHT algorithm. These recognition data for the sketches sounds are compared to an upper baseline: the average recognition performance for the original sounds (black line). They are also compared to a lower baseline: the chance level, i.e. 25% here (dotted gray line).

It also showed that performance was overall better for the PP algorithm than for the IHT algorithm $[F(1,9) = 54.72; p < 0.0001]$. All second-order interactions were also significant:
[features $\times$ algorithm : $F(2,18) = 85.92; p < 0.0001$.
features $\times$ noise : $F(2,18) = 32.39; p < 0.0001$.
algorithm $\times$ noise : $F(1,9) = 49.46;\ \ p < 0.0001$]. Finally, the third-order interaction was also significant $[F(1,9) = 28.07; p < 0.0001]$, and highlighted that the only exception for which the IHT algorithm outperformed the PP algorithm was in the noise condition, with 1000 feat/s.

4 Discussion

The main aim of this study was to investigate the feasibility of the auditory sketches idea. From the results, it seems that the sketches design method outlined in Fig. 1 has some potential. In the experiments, even though the vast majority of the parameters was omitted, the perceptual task (emotion recognition for nonverbal interjections) was performed well above chance: sketches retained some of the relevant information with as little as 10 features/seconds. More information-theoritic work remains to be done on quantifying the sparsity

that was actually achieved, because features/second is an imperfect measure, but the results nevertheless strongly suggest that sparse representations of sounds based on biologically-motivated models produce perceptually relevant results.

Further observations can be made by comparing the variants we tested for the sketches process. Perhaps surprisingly, a state-of-the-art sparse decomposition algorithm minimizing reconstruction error (IHT) did not lead to better results than a simple peak-picking and thresholding (PP) without any optimization. In fact, in general, the reverse was true, and PP largely outperformed IHT. These preliminary results need to be extended with a larger variety of stimuli and perceptual tasks, but still, we can speculate on such an outcome. Because auditory models are inspired by the physiology of the human hearing system, they may be particularly relevant as an auditory representation. A simple algorithm like PP, although not optimal (in the least-square sense for the approximation), may be enough to capture important features by sampling some of the important landmarks of the representations.

Fig. 6 illustrates this point, by highlighting an important difference between the two selection algorithms. The PP algorithm tends to select relatively distant atoms (see Fig. 6(a)) as an extended high-energy patch in the representation can be summarized with a single peak. In contrast, the IHT algorithm will attempt to capture accurately such high-energy patches and will use several atoms to do so (see Fig. 6(c)). These opposite behaviors lead to different reconstructions: whereas IHT achieves a highly precise reconstruction of some particular parts of the original spectrogram (see Fig. 2 and Fig. 6(d)), this is done at the expense of smaller coverage of the whole parameter space.

However, we should point out that it is probably too early to generalize the superiority of a local maxima detection over a least-squares approach. The IHT algorithm constitutes one possible way to solve problem (4) amongst a large number of possibilities. We chose IHT for implementation and complexity reasons (see Subsect. 2.2) but other algorithms could potentially improve the results (see *e.g.*, approaches based on a problem relaxation [21,22], or greedy algorithms [23]). The *sparsity-at-analysis* point of view can also be questioned, and could be compared to more standard synthesis approaches. Further experiments could finally investigate some other sparsifying procedures, intermediate between peak-picking and energy-maximizing, for instance iterative procedures based on a time-frequency masking model [24].

Another surprising result is the overall better performance for the auditory spectrogram representation compared to the cortical one. One of the limitations of the sounds we used was their short duration (around 1s). The cortical model contains filters tuned to longer modulations, so it is possible that any potential benefit of the spectro-temporal analysis only becomes apparent for longer sounds.

Finally, we found that the recognition of sketches was robust to a moderate amount of noise, but less so than for the original signal. This is in line with many psychophysical observations showing that degraded signals are more susceptible to noise. Nevertheless, one hypothesis for the sketches was that sparsification would lead to some denoising. Our results suggest that either the representations

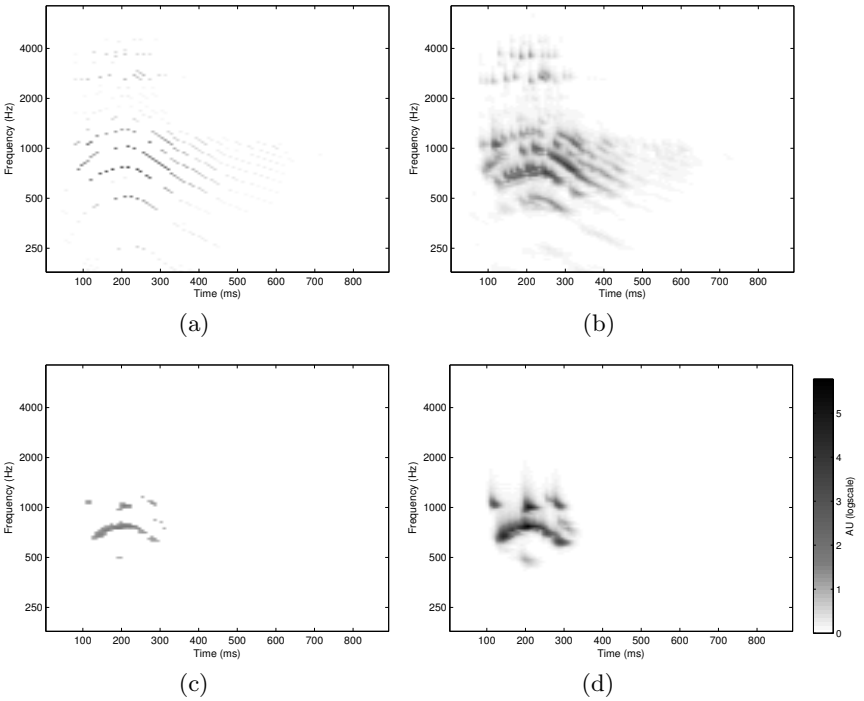

Fig. 6. Sparse auditory spectrograms obtained by means of the PP algorithm ((a) and (b)) and the IHT algorithm ((c) and (d)), directly after the decomposition ((a) and (c)) and after resynthesis of the audio signal ((b) and (d)). Here, we keep 100 feat./s. AU: arbitrary units, log scale.

failed at this goal, or that, more likely, the selection procedure could be improved. Such an approach has proven successful for denoising of speech signals, with the cortical model [25]: by increasing the dimensionality of the representation, noise and signal get mapped into different parts of the parameter space.

5 Perspectives

This preliminary study already shows that only a few features extracted from an auditory-based representation can produce a sound with recognizable perceptual traits. Even though the resulting sketch may be highly distorted compared to the original, under certain constraints, the selected features can be sufficient for recognition of complex properties such as emotional content. Obviously, more work remains to be done on each stage of the sketching process, and in particular, the iterative nature of the algorithm needs to be put to the test.

In addition, a few ideas emerge on how sound features should be combined in order to build recognizable auditory sketches. For a task of sound recognition, it seems that it is more important to have some cues on how energy is spread in

the time-frequency plane, rather than a precise description of the most energetic components. Interestingly, this is similar to what is being done in state-of-the-art audio fingerprinting techniques, that choose salient points as local maxima in large blocks on the time-frequency plane. More precisely, it seems that the right way to select atoms is not purely based on energy criteria, but also their information content: we need to select a set of atoms that carry energy but also whose mutual information is minimal. In other words, we shift from the standard paradigm of sparsity justified by Occam's razor (amongst 2 explanations, prefer the one that is simplest) to an "informed" version (amongst 2 explanations, prefer the one that brings you more information on top of a prior model). This brings us close to the original sketches metaphor: to sketch a visual object, an artist will usually not attempt photographic realism. Rather, in a few pencil lines, an attempt will be made to capture what makes this object unique. It is our hypothesis that such an approach may have interesting implications for signal processing, but also for understanding how human listeners perform recognition tasks (see e.g. [26]).

Acknowledgements. This work was partly funded by the Fondation Pierre-Gilles de Gennes pour la Recherche. Laurent Daudet is on a joint position with Institut Universitaire de France. We thank Florence Bouhali for collecting the psychophysical data. We also thank Shihab Shamma and Nima Mesgarani for their input at various stages of the project. Some of these ideas were developed in the Telluride Neuromorphic Engineering workshop.

References

1. Mallat, S.: A Wavelet Tour of Signal Processing - The Sparse Way, 3rd edn. Academic Press (December 2008)
2. Gabor, D.: Acoustical quanta and the theory of hearing. Nature 159, 591–594 (1947)
3. Plumbley, M., Blumensath, T., Daudet, L., Gribonval, R., Davies, M.: Sparse representations in audio and music: From coding to source separation. Proceedings of IEEE 98(6), 995–1005 (2010)
4. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer (2010)
5. Aharon, M., Elad, M., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. on Signal Processing 54(11), 4311–4322 (2006)
6. Shannon, R., Zeng, F., Kamath, V., Wygonski, J., Ekelid, M.: Speech recognition with primarily temporal cues. Science 270(5234), 303–304 (1995)
7. Patterson, R., Allerhand, M., Giguére, C.: Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. Journal of the Acoustical Society of America 98(4), 1890–1894 (1995)
8. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. Journal of the Acoustical Society of America 118(2), 887–906 (2005)
9. Patil, K., Pressnitzer, D., Shamma, S., Elhilali, M.: Music in our ears: the biological bases of musical timbre perception. PLoS Comp. Biol. 8(11), e1002759 (2012)

10. Portilla, J.: Image restoration through l0 analysis-based sparse optimization in tight frames. In: Proc. IEEE Int'l Conference on Image Processing (ICIP), pp. 3865–3868 (2009)
11. Belin, P., Fillion-Bilosdeau, S., Gosselin, F.: The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods 40(2), 531–539 (2008)
12. Elhiliali, M., Chi, T., Shamma, S.A.: A spectro-temporal modulatio index (stmi) for assessment of speech intelligibility. Speech Communication 41(2-3), 331–348 (2003)
13. Griffin, D., Lim, J.: Signal reconstruction from short-time fourier transform magnitude. IEEE Trans. Acoust., Speech, and Signal Proc. 32(2), 236–243 (1984)
14. Sturmel, N., Daudet, L.: Signal reconstruction from its STFT magnitude: a state of the art. In: Proc. International Conference on Digital Audio Effects, DAFx 2011 (2011)
15. Yang, X., Wang, K., Shamma, S.A.: Auditory representations of acoustic signals. IEEE Trans. on Information Theory 38(2), 824–839 (1992)
16. Drémeau, A., Herzet, C., Daudet, L.: Boltzmann machine and mean-field approximation for structured sparse decompositions. IEEE Trans. on Signal Processing 60(7), 3425–3438 (2012)
17. Elad, M., Milanfar, P., Rubinstein, R.: Analysis versus synthesis in signal priors. Inverse problems 23(3), 947–968 (2007)
18. Blumensath, T., Davies, M.E.: Iterative thresholding for sparse approximations. Journal of Fourier Analysis and Applications 14(5-6), 629–654 (2008)
19. Hoogenboom, R., Lew, M.: Face detection using local maxima. In: Proc. Int'l Conference on Automatic Face and Gesture Recognition, 334–339 (1996)
20. Schwartzman, A., Gavrilov, Y., Adler, R.J.: Multiple testing of local maxima for detection of peaks in 1d. Annals of Statistics 39(6), 3290–3319 (2011)
21. Chambolle, A.: An algorithm for total variation minimization and application. Journal of Mathematical Imaging and Vision 20(1-2), 89–97 (2004)
22. Peyré, G., Fadili, J.: Learning analysis sparsity priors. In: Int'l Conference on Sampling Theory and Applications, SAMPTA (2011)
23. Nam, S., Davies, M., Elad, M., Gribonval, R.: Cosparse analysis modeling - uniqueness and algorithms. In: Proc. IEEE Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5804–5807 (2011)
24. Balazs, P., Laback, B., Eckel, G., Deutsch, W.: Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking. IEEE Transactions on Audio, Speech and Language Processing 18(1), 34–39 (2010)
25. Mesgarani, N., Shamma, S.A.: Speech enhancement using spectro-temporal modulations. EURASIP Journal on Audio, Speech, and Music Processing V, ID 42357 (2007)
26. Agus, T.A., Suied, C., Thorpe, S.J., Pressnitzer, D.: Fast recognition of musical sounds based on timbre. Journal of the Acoustical Society of America 131(5), 4124–4133 (2012)