SOUND ENHANCEMENT USING SPARSE APPROXIMATION WITH SPECLETS

Manuel Moussallam, Pierre Leveau, Si Mohamed Aziz Sbaï

Audionamix
204, rue de Crimée, 75020 Paris, France
firstname.lastname@audionamix.com

ABSTRACT

This paper addresses an innovative approach to informed enhancement of damaged sound. It uses sparse approximations with a learned dictionary of atoms modeling the main components of the undamaged source spectra. The decomposition process aims at finding which of the atoms could constitute the decomposition of the undamaged source in order to recover it. The decomposition of the damaged signal is done with a Matching Pursuit algorithm and involves an adaptation of the dictionary learned on undamaged sources. Evaluation is performed on a bandwidth extension task for various classes of signals.

***Index Terms*—** sparse representations, Matching Pursuit, audio signal enhancement, dictionary learning

1. INTRODUCTION

Sound enhancement consists in modifying a signal in order to improve its perceived quality or its fidelity to an audio source category. Recent studies have proposed generic algorithms with regard to the sources and for well-calibrated sound defects, some others are focused on a specific sound source.

It includes a large amount of principles, goals and techniques that are not necessarily comparable one to another. However, the problem often comes down to trying to recreate missing parts of the signal's spectrum, usually high frequencies.

An obvious distinction between sound enhancement approaches is the use or not of prior knowledge on the processed signal. When no such information is available, it deals with *blind* enhancement. Scientific literature on the subject is abundant since this problem has many applications in the communication area, especially for voice signals. For bandwidth extension tasks, techniques have been developed both in time and frequency domain, with recent advances in the latter.

Time domain methods include non linear distortions of signals. A trivial example is taking the square or the absolute value of each sample then subtract the arithmetic mean and normalize. These methods are very cheap but produce terrible results when dealing with harmonic signals, for they create very disturbing intermodulation.

Frequency Translation methods, also known as Band Replication, works in the frequency domain. The Spectral Band Replication (SBR) process is the most commonly used, since it has been adopted in the MPEG-4 HE-AAC norm[1]. SBR works at the decoder side and uses extra information transmitted alongside the data to reconstruct the high-frequency parts of the signal. Several improvements have since been made in [2, 3].

However, harmonic parts of the spectrum are generally poorly enhanced with these methods in blind context, especially when dealing with polyphony. Dictionary-based methods such as [4] present an innovative approach to help tackle these limitations. Recently developed object representations then provides an interesting framework for this application, since they work on high level, time-frequency structures.

In this paper, a new method for sound enhancement is proposed. It aims at being generic for sources, provided it has been trained on undamaged samples of the source, and applicable to stationary defects in the frequency contents. It relies on the learning of dictionaries that model undamaged source spectra, whose supposed activation is detected in the signal to enhance. The chosen framework is the domain of sparse approximations of signals, which provides the possibility to decompose the sound into an object representation.

The enhancement algorithm relies on several steps :

1. Learning of the dictionary on undamaged sources ;
2. Decomposition of the signal to enhance with the adapted learned dictionary into an object representation ;
3. Post-processing of the object representation ;
4. Synthesis of the processed object representation .

The object representation process is close to the one proposed in[5]. An overview is presented in Section 2. The post-processing of the object representation is shown in Section 3 and the results it provides in Section 4.

2. OBJECT REPRESENTATION

2.1. Signal model: Speclets

2.1.1. *Definition*

The original signal is approximated by a linear combination of atoms learned on the undamaged source, α_λ taking values in $\mathbb{C}$:

$$x \simeq \sum_{\lambda \in \Lambda} \alpha_\lambda s_\lambda \tag{1}$$

The atoms s_λ, that we name *speclets*, are build themselves with Gabor atoms. A Gabor atom is defined by:

$$g_{s,u,f}(t) = \frac{1}{\sqrt{s}}\, w\left(\frac{t-u}{s}\right) e^{2j\pi(f(t-u))} \tag{2}$$

where s is the scale of the atom, u its time localization, f its frequency, and w a Gaussian window with a unitary energy.

The speclets $s_{\lambda=(s,u,F,A,\Phi)}$ are linear combinations of Gabor atoms that have the same scale and time support, but different fre-

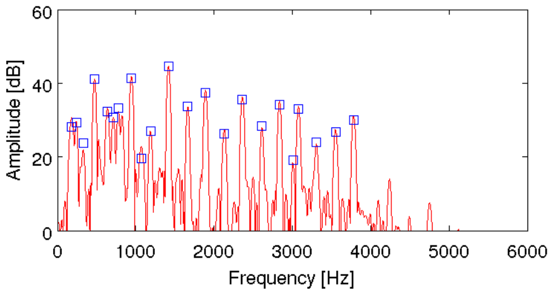

Fig. 1. *Modulus of the Fourier Transform of an analyzed signal (continuous line), and speclet modeling this signal. The speclet's components parameters are (frequency, amplitude) couples (squares), and match the peaks of the spectrum of the analyzed signal.*

quency localizations:

$$s_{s,u,F,A,\Phi} := \sum_{m=1}^{M} a_m e^{j\Phi_m} g_{s,u,f_m}, \text{ with } |\langle g_{s,u,f_i}, g_{s,u,f_j}\rangle| < \epsilon \tag{3}$$

for $i \neq j$, where $F = (f_m)_{m=1..M}$ is the vector of frequency localizations of the Gabor atoms, $A = (a_m)_{m=1..M}$ their respective weights and $\Phi = (\phi_m)_{m=1..M}$ their respective phases. The A vector is normalized so that the speclet has a unit energy. In the speclet atom definition (3), a constraint is set on the orthogonality of the Gabor atoms, which has to be limited to a value ϵ. This constraint has an interest for the signal decomposition step, that will be highlighted in section 2.2.2.

Speclets capture a major part of the timbre information of acoustic sources: its goal is to model the spectrum of a single time frame of a source with a few Gabor atoms. Figure 1 displays a speclet along with the spectrum it models: the frequencies of the Gabor atoms are aligned with the highest peaks of the spectrum. Defining atoms that capture such timbre characteristics is not new: Gribonval [6] defined harmonic atoms that are extracted in a unsupervised way (without learning) but with given frequency localizations, using a modified Matching Pursuit algorithm. This idea has been extended to learned harmonic atoms in [5]. Cho [4] also used high-level atoms composed of Gabor atoms, with arbitrary scales, time localizations and frequency localizations, but without orthogonality assumption for the components taken pairwise, preventing to compute the projections in the spectral domain.

2.1.2. Learning

The learning of the speclet dictionary is not detailed here for the sake of conciseness. It is based on a Matching Pursuit algorithm with Gabor atoms, which is performed on signal frames of the undamaged sources as in [4]. To reduce the number of learned speclets, a vector quantization algorithm is perfomed the whole speclet dictionary. Before the quantization, the spectra of the speclet atoms are gaussianized to reduce the distance between speclet that have a small fundamental frequency difference.

2.2. Decomposition

The decomposition algorithm provides a sound object representation of the signal using the aforementioned signal model.

2.2.1. Matching Pursuit algorithm

The Matching Pursuit (MP) algorithm [7, 8] is a popular algorithm for sparse approximations of signals. It provides a generally suboptimal sparse representation of the analyzed signal, but at a reasonable computational cost. It consists in the following steps:

```
1. Compute the inner products between the
   signal and the atoms in the dictionary
   ($|\langle x, s_\lambda \rangle|$);
2. Select the atom in the dictionary that
   maximizes the modulus of its inner
   product with the signal:
   $s_{\lambda_0} := \arg\max_{s_\lambda \in \mathcal{D}} \{|\langle x, s_\lambda \rangle|\}$;
3. Subtract the atom with its weight from
   the signal: $x \leftarrow x - \langle x, s_{\lambda_0} \rangle . s_{\lambda_0}$ . Go to
   step 1;
```

If the algorithm is run on an entire signal, the update of the inner products is only performed on the timezone where lay the extracted atom (as in [9]).

The analysis of a real signal with complex atoms remains valid by performing the pursuit with couples of conjugate atoms, as mentioned in a number of studies, e.g. in [7]. For Gabor atoms, the real atom is a windowed cosine, whose phase can be computed as follows [7]:

$$e^{j\phi} = \frac{\langle x, g_{s,u,f} \rangle}{|\langle x, g_{s,u,f} \rangle|} \tag{4}$$

2.2.2. Matching Pursuit with speclets

To deal with specific types of atoms, MP is often adapted. For the speclet case, the adaptation lies in the computation of the inner products. The modulus of the inner product of the signal with the speclet atom writes as follows:

$$|\langle x, s_\lambda \rangle| = \left| \sum_{m=1..M} a_m e^{j\phi_m} \langle x, g_{s,u,f_m} \rangle \right| \tag{5}$$

The phases of each component is computed as in (4): they are stated as being adapted to the signal. This computation of component phases is only valid if the Gabor components are orthogonal enough. If this assumption is not valid, the Matching Pursuit can diverge. As a consequence, the angle of $e^{j\phi_m}\langle x, g_{s,u,f_m} \rangle$ is 0. The inner product can thus be written:

$$|\langle x, s_\lambda \rangle| = \sum_{m=1..M} a_m |\langle x, g_{s,u,f_m} \rangle| \tag{6}$$

The modulus of the inner product between a speclet and the signal can thus be seen as inner product of an amplitude vector and some points of the modulus of a Fourier transform, which considerably reduces the computation time of the step 1 of the MP algorithm.

With the help of this algorithm, a signal can be decomposed as a resynthesizable object representation on which post-processing in the object domain can be applied. Each object is an atom whose spectral shape has been learned on an audio source.

2.2.3. Molecular Matching Pursuit with speclets

An extension of MP has been recently proposed in [8], where higher-level structures called *molecules* are defined. They are constituted of a set of atoms with common properties (such as timbre-, time-proximity) and can capture high-level musical information such as

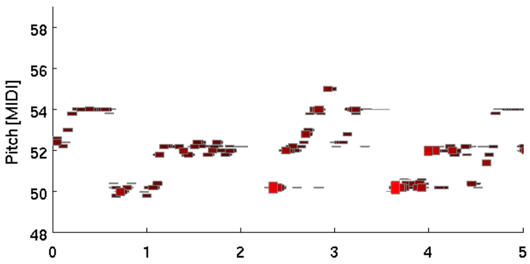

Fig. 2. *Object representation in the time-pitch plane of a clarinet solo using clarinet atoms. Atoms are rectangles whose length represent the atom durations and height their amplitudes.*

notes. Molecules are obtained with a Viterbi algorithm (not detailed here) working on a grid obtained by projecting of the signal on the speclet dictionary. The transition costs are defined to reflect timbre proximity or semantic constraints, such as brutal note changes. The signal m related to a molecule $\mathcal{M}$ can be written as follows:

$$m = \sum_{l \in \Lambda_{\mathcal{M}}} p_l . s_l \tag{7}$$

The atoms $\{s_l\}$ composing the molecule $\mathcal{M}$ are temporally adjacent and overlapping. MP algorithm is then adapted as follows:

```
1. Compute the inner products between
   the signal $x$ and the atoms in the
   dictionary.  A grid $G$ is thus obtained;
2. Find the optimum path through $G$ with a
   Viterbi algorithm using a transition
   weights matrix $T_{\mathcal{D}}$:
   $\mathcal{M} = Viterbi(G, T_{\mathcal{D}})$
3. Orthogonal projection of the molecule to
   recalculate atom weights :
   $\mathbf{p} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T x$ .Subtract the molecule
   from signal $x \leftarrow x - \mathbf{Mp}$ then go to step 1;
```

where $\mathbf{p}$ is the vector of the speclet coefficients in the molecule p_l and $\mathbf{M}$ the matrix of the speclet waveforms.

Molecules efficiently capture the inherent consistency of musical notes, thus providing a robust support for sound enhancement in a polyphonic context. Figure 2 depicts an example of object representation of a solo clarinet in the time-pitch plane.

3. POST-PROCESSING

An object representation of the signal is obtained once the decomposition step is over: spectral enhancement can then be performed. The whole process is depicted on Figure 3. Let x_f be a version of an original signal x with stationary damaging mask U and additional noise w.

$$x_f = U.x + w \tag{8}$$

Finding the correct x when observing x_f is solving an inverse problem for sound enhancement. Assuming that we have learned a undamaged source-adapted dictionary D on a set of full band signals, then we might be able to replace speclets from the decomposition of x_f by full-band versions of the same speclets. The missing frequencies would thus be recovered with very high consistency. A simple MP-decomposition of x_f on D would not provide good results, because the chosen speclets would have an insufficient and biased correlation with x_f. As a consequence, some speclets would be chosen to compensate the high frequency by having their Gabor phases opposing them of the first speclet. The one-to-one matching of the speclet and the source would be lost, and the signal could not be enhanced. To adapt the dictionary to the signal to enhance, a stationary mask $\tilde{U}$ transforming the global envelope of D into the global envelope of x_f is either guessed or estimated, and applied to D, yielding a filtered dictionary D_f. A MP decomposition with D_f is then performed to decompose x_f, leading to:

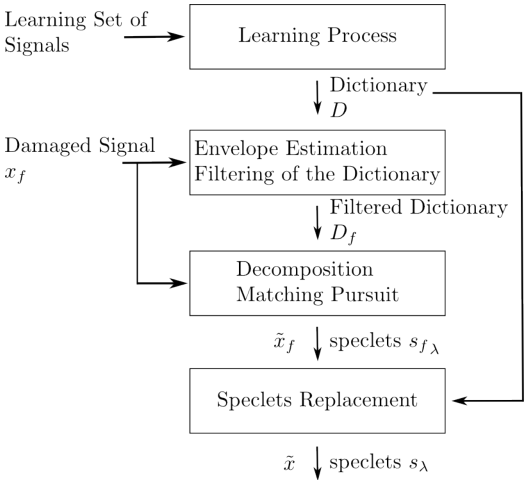

Fig. 3. *Overview of the enhancement process*

$$\tilde{x}_f = \sum_{\lambda \in \Lambda} \alpha_\lambda s_{f_\lambda} \tag{9}$$

where the s_{f_λ} are speclets from D_f. The post-processing step consists in the replacement of these speclets by their corresponding full-band versions in D. The result is an enhanced version of the reconstituted signal, with recreated missing frequencies.

$$\tilde{x} = \sum_{\lambda \in \Lambda} \alpha_\lambda s_\lambda \tag{10}$$

Unfortunately, the phases of this recovered frequencies cannot be calculated using the signal, since it does not contain these frequencies. Phase consistency can however be forced between consecutive atoms. Working with molecules facilitates this process, for they capture higher level structures such as musical notes. The phase recovery processes as follows: for overlapping atoms, phases of Gabor components ϕ_{f_1,t_1} and ϕ_{f_2,t_2} are adapted so that the sum is optimized at the maximum overlapping point t_M by:

$$\phi_{f_2,t_2} = \phi_{f_1,t_1} + 2\pi(f_2 t_2 - f_1 t_1 + (f_1 - f_2)t_M) \tag{11}$$

Ensuring phase consistency between atoms in the same molecule provides good perceptive results.

4. EXPERIMENTS AND RESULTS

Experiments have been conducted on a bandwidth extension task, where the damaging mask U is a destructive low-pass filtering. Evaluation is performed on synthetic harmonic series, both monophonic

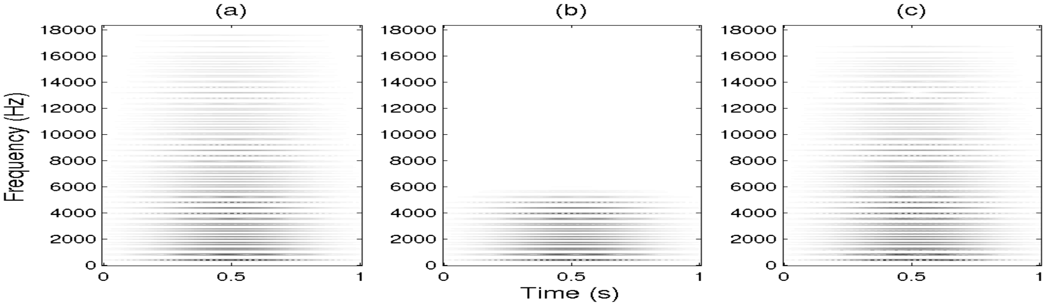

Fig. 4. *Spectrograms of (a) original polyphonic harmonic serials of 40 partials, f0 = 440 Hz and 512 Hz, (b) damaged signal cutoff frequency of 6000 Hz and (c) restored signal. Almost perfect signal reconstitution is achieved*

and polyphonic, then on real musical data, i.e. short instrument solo segments and mixtures. Decompositions were performed with 200 speclets per second of signal, speclet length is 93 ms while hop size is 23 ms, full-band speclets contain 40 Gabor atoms and FFT size is 4096. The learning process is conducted on similar harmonic series for synthetic signals, and on RWC database for instrument solos. A few atoms are kept for each pitch, instrument and style. The more prior knowledge on the sources, the smaller the dictionary. In blind contexts, typical dictionary size is 10000 atoms, but in single instrument contexts, this can be reduced to a few hundreds.

Then the technique is applied to enhance an artificially damaged segment [1].

An objective evaluation criteria is defined in the frequency domain. Indeed, since high frequency phases are artificially recreated, it is uneasy to compare original and restored signals in the time domain. Perceptively similar signals might have poor sample correlation, however their spectrograms are very close. The following High-Frequency Gain (HFG) for a signal $\tilde{x}$ whose spectrogram is written $\tilde{X}(f,t)$ is defined:

$$HFG = 10.\log\left(\frac{\|\tilde{X}(f,t)\|^2}{\|\tilde{X}(f,t) - X(f,t)\|^2}\right) \qquad (12)$$

Signal	HFG
Synthetic Mono	35.2
Synthetic duo	34.8
Trumpet solo (1 sec)	18.2
Trumpet solo (10 sec)	13.1
Clarinet solo (10 sec)	9.9
Violin + piano	7.5
Clarinet + mixture	6.3

Table 1. *Enhancement results (in dB) for different classes of signals, damaged signals are destructively low-pass filtered at 6000 Hz cutoff frequency. The HF gain column show gain in the damaged parts or in the restored area*

Table 1 depicts the results of the enhancement process. For synthetic signals, the reconstruction is nearly perfect for both solo and polyphonic cases. For solo instruments, results highly depend on the dictionary used. For sound mixtures, only the soloing instrument is enhanced, objective evaluation is then very uneasy to perform. However perceptive results are encouraging.

5. CONCLUSION AND FUTURE WORK

In this study, the use of sparse approximation of the signal with dictionaries learned on undamaged signals has been investigated for sound enhancement. It enables the recovery of missing frequencies. Improvements can be conducted at the decomposition level by raising its ability to extract relevant sound objects. Post-processing could also be improved by adding constraints on the time evolution of the amplitude of the note partials. A subjective quality assessment of this algorithm will also be performed in future work.

6. REFERENCES

[1] M. Wolters, K. Kjrling, D. Homm, and H. Purnhagen, “A closer look into mpeg-4 high efficiency aac,” In 115th AES Convention, Tech. Rep.

[2] A. J. S. Ferreira and S. Deepen, “Accurate spectral replacement,” in *AES 118th Convention, Paper Number: 6383*, 2005.

[3] T. Zernicki and M. Bartkowiak, “Audio bandwidth extension by frequency scaling of sinusoidal partials,” in *AES 125th Convention, Paper Number: 7622*, 2008.

[4] N. Cho, Y. Shiu, and C. Kuo, “Efficient music representation with content adaptive dictionaries,” in *Proc. of Int. Symp. on Circuits and Systems (ISCAS)*, 2008, pp. 3254–3257.

[5] P. Leveau, E. Vincent, G. Richard, and L. Daudet, “Instrument-Specific Harmonic Atoms for Mid-Level Music Representation,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, p. 116, 2008.

[6] R. Gribonval and E. Bacry, “Harmonic decomposition of audio signals with matching pursuit,” *IEEE Trans. on Signal Processing*, vol. 51, no. 1, pp. 101–111, January 2003.

[7] S. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[8] L. Daudet, “Sparse and structured decompositions of signals with the molecular matching pursuit,” *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 14, pp. 1808–1826, 2006.

[9] S. Krstulovic and R. Gribonval, “Mptk: matching pursuit made tractable,” *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, May 2006.

[1] Available [temporary address for review purposes] at http://research.audionamix.com/private/45/63080091641fe62902a39f19f3613a2c/