How Sparsely Can a Signal be Approximated while Keeping its Class Identity?

Manuel Moussallam , Thomas Fillon[*] ,
Gaël Richard
Institut Telecom - Telecom ParisTech -
CNRS/LTCI
37, rue Dareau
75014 Paris, France
first.last@telecom-paristech.fr

Laurent Daudet
Institut Langevin - UMR 7587
Univ. Paris 7 Diderot - ESPCI ParisTech
10, rue Vauquelin
75005 Paris, France
laurent.daudet@espci.fr

ABSTRACT

This paper explores the degree of sparsity of a signal approximation that can be reached while ensuring that a sufficient amount of information is retained, so that its main characteristics remains. Here, sparse approximations are obtained by decomposing the signals on an overcomplete dictionary of multiscale time-frequency "atoms". The resulting representation is highly dependent on the choice of dictionary, decomposition algorithm and depth of the decomposition. The class identity is measured by indirect means as the speech/music discrimination power of features derived from the sparse representation compared to those of classical PCM-based features. Evaluation is performed on French Broadcast TV and Radio recordings from the QUAERO project database with two different statistical classifiers.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Signal analysis, synthesis, and processing*

General Terms

Algorithms,Experimentation

1. INTRODUCTION

For some time now, sparse decomposition algorithms have appeared as powerful tools for compactly representing a signal. The main idea is to find amongst a number of predefined waveforms (dictionary), a small (sparse) linear combination of them that efficiently captures most of the signal information. Given a dictionary $\mathcal{D}$, a discrete signal $x[n]$ in $\mathbb{R}^N$ can be written:

$$x[n] = \sum_{\lambda \in \Lambda} \alpha_\lambda \Phi_\lambda[n] + \epsilon[n] \tag{1}$$

where Λ is the set of parameters for the non-zero encoding coefficients, with a cardinality hopefully much smaller than N, Φ_λ are the so-called *atoms* from the dictionary and $\epsilon[n]$ is the residual. While there exist different strategies, it is commonly acknowledged that not only do sparse representations allow coding gain improvements, but also higher-level tasks such as source separation [3] or sound enhancement. Sparsity arises from the overcompleteness of the dictionary and convergence properties of decomposition algorithms such as *Matching Pursuit* [4]. Such algorithms are the results of extensive research for tackling the combinatorial problem of minimizing both the size of Λ (i.e., the sparsity) and the residual's energy (i.e., the agreement between model and data). However, little has been done so far to evaluate how the exact degree of sparsity influences the high-level information content of the representation, in other words how far sounds can be simplified while retaining their "identity".

This ill-posed problem is here addressed using Ravelli's framework [6] - where overcompleteness is the result of cumulating *Modified Discrete Cosine Transforms* (MDCT) of different scales - since he demonstrated in [7] the usefulness of his codec for indexing purposes. The adequacy of the representation for indexing purposes will be indirectly measured as performance in a simple speech/music discrimination task.

The *speech / music* classification of digital audio streams has been extensively examined over the past decades, both on compressed [2, 10] and uncompressed data [8, 9, 5] . A wide range of applications such as *Automatic Speech Recognition* systems, *multi-modal coding* schemes, indexing, segmentation and information retrieval of audio data relies on such systems. Most proposed methods share the same global pattern: audio stream is sliced in temporal frames, then a range of features is computed on each frame. Temporal integration and statistical classification algorithms finally yield a speech/music segmentation of the audio stream.

Given the simultaneous time-frequency resolution available with a union of MDCT bases, it is possible to derive low-level features from the sparse approximations and evaluate their performances for a given task. A study of spar-

[*]This work was partly supported by the QUAERO Programme, funded by OSEO, French State agency for innovation

sity influence can then be done. The important point is that we want to take advantage of the intrinsic simplicity of the sparse representation, and the meaningfulness of its parameters, by computing these features directly in the sparse domain, instead of resynthesizing the signal and computing the standard features. As we shall see, the computation of features is here always extremely simple.

This article goes as follows: Section 2 recalls Ravelli's framework of sparse representation, Section 3 describes the new proposed features and classification schemes. In Section 4 a search for sufficient sparsity levels is conducted then comparison is made with classical features over the QUAERO dataset, then Section 5 suggests some conclusions.

2. DECOMPOSITION IN A UNION OF MDCT BASES

Solving equation 1 while minimizing both the size of Λ and the residual's energy is a combinatorially hard problem. Greedy algorithms such as *Matching Pursuit* [4] usually provide suboptimal solutions at a reasonable cost. Sparsity can often be enforced by designing appropriate overcomplete dictionaries. Ravelli [6] proposes to use a union of M MDCT bases (called blocks), at 8 different dyadic scales (i.e., frame size), to build an overcomplete dictionary $\mathcal{D} = \bigcup_{m=0}^{M-1} \mathcal{D}_m$:

Atoms from block m all have the same length L_m and have the following form:

$$\Phi_{m,p,k}[n] = w_m[u]\sqrt{\frac{2}{L_m}}\cos\left[\frac{\pi}{L_m}\left(u+\frac{L_m+1}{2}\right)\left(k+\frac{1}{2}\right)\right] \quad (2)$$

and $u = n - p.L_m - T_m$, where p denotes the frame index, k the frequency bin index, and T_m a temporal offset needed to align different window sizes. In the following we will use the notation i_m to summarize parameters m,p and k. Each MDCT block is in itself an orthogonal base but combining them brings overcompleteness. Decomposing on such a dictionary yields a representation of the form:

$$\tilde{x}[n] = \sum_{m=0}^{M-1}\sum_{i_m=0}^{I_m-1} \alpha_{i_m}\Phi_{i_m}[n] \quad (3)$$

where I_m is the number of atoms picked from block m to represent the signal. It is worth noticing that because the *Matching Pursuit* algorithm substracts each selected atom from the signal, it is not equivalent as to performing M MDCT in parallel and selecting the largest coefficients in each one.

The depth of the decomposition is expressed by the strictly increasing *Signal-To-Residual Ratio* (SRR) defined by:

$$SRR = 10\log\left(\frac{\|\sum_{\lambda\in\Lambda}\alpha_\lambda\Phi_\lambda[n]\|_2^2}{\|\epsilon[n]\|_2^2}\right) \quad (4)$$

An exemple of pseudo-sepctrogram of such decomposition on a short monophonic trumpet signal is given figure 1.

3. SPEECH/MUSIC DISCRIMINATION

3.1 New Features

The decomposition in equation 3 can be easily retrieved from coding coefficients [6]. This multi-scale framework allows a great precision both in time and frequency. It is intuitive to reckon that energy repartition in the different scales as well as the number of coefficients needed to reach a given SRR should depend on the class of data. Two very simple scale-related features can then be derived:

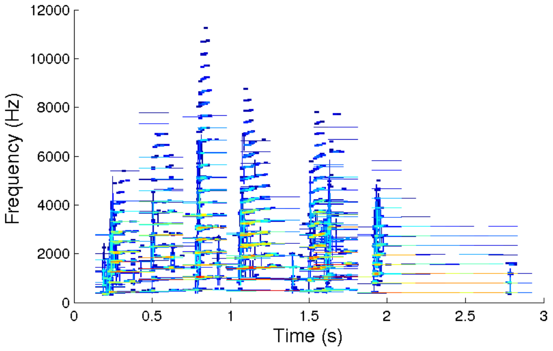

Figure 1: ***Multi-scale Time Frequency sparse representation of a short trumpet signal using 3 MDCT bases of scale 128, 1024 and 8192 samples and a SRR of 20 dB in the decomposition. Transient parts are represented by small scale MDCT atoms (narrow in time) while sustained harmonics are represented by longer ones (thin in frequency).***

Scale Coefficient Count (SCC): is simply the vector of I_m or the number of atoms selected on each MDCT block.

Scale Amplitude Repartition (SAR): normalized vector of the summed amplitudes of atoms selected on each block:

$$SAR_m = \frac{\sum_{i_m=0}^{I_m-1}|\alpha_{i_m}|}{\sum_{m=0}^{M-1}\sum_{i_m=0}^{I_m-1}|\alpha_{i_m}|} \quad (5)$$

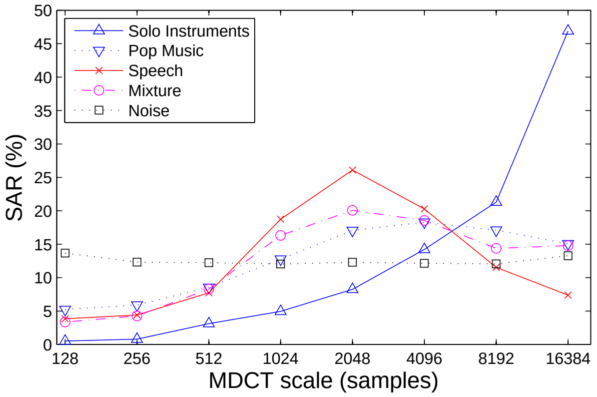

Figure 2: ***Centroids of SAR features at 10 dB using 8 MDCT bases with L_m ranging from 128 to 16384.***

Figure 2 shows the SAR vector centroids for different classes of signals at 10 dB of SRR. Solo instruments are taken from the RWC database. Pop music, Speech and Mixture signals are taken from radio broadcast recordings. Speech and Solo Instruments can be easily distinguished, but Mixture and Pop music present fewer differences. Above 5 dB, the SRR yields little modification to these profiles.

Table 1 shows the average number of atoms per second for the same signals and various decomposition depths. With a fixed decomposition depth the number of coefficients (and so bitrates) needed to represent monophonic musical signals is far less than for speech and again less than for pop music. SCC appears to provide useful independent additional information to discriminate between these classes.

	SRR (dB)				
Class	**1**	**5**	**10**	**15**	**20**
Pop Music	3	48	314	926	1893
Solo Instrument	1	6	21	61	190
Speech	2	21	89	267	625
Mixture	2	24	116	372	883
Noise	324	1969	4145	6302	8419

Table 1: ***Average number of atoms per second, for various SRR and classes of signal.***

3.2 Classification techniques

Different types of classifiers can be used to perform a speech / music discrimination task. Linear Discriminant Analysis (LDA) is a powerful statistical tool for such problems. It finds a linear combination of the features that best satisfies a naive bayes distribution assumption. Abundant litterature is available on the subject and efficient implementations are easily found on the web. Here, the implementation from Matlab©Statistics Toolbox^{TM}is used. Support Vector Machines (SVM) have been thoroughly used by the machine learning community. This kernel-based method finds a hyperplan that separates the annotated training data set in classes with a maximum margin. In this paper, the implementation described in [1] is considered. Both classifier are used to evaluate the merit of the proposed features, compared to the classical MFCC and Δ-MFCC features.

4. EXPERIMENTS AND RESULTS

4.1 Finding the right decomposition depth

To begin with, evaluation is performed on one hour of radio broadcast (sliced in 10 parts of 6 minutes each) taken from the QUAERO project database, composed of popular western musical pieces, news report and interviews. Features are calculated for several depths of decomposition (SRR) on frames of size 16384 samples with 50 % overlap. Then a LDA classifier is trained on 8 random slices and tested on its classification output of the other 2. To cross-validate the results, 10 random permutations are taken. Figure 3 shows mean error results for speech and music detection tasks with the SAR features and the combination of SAR and SCC. It appears that increasing the SRR leads to better results, however it also means less sparsity.

To evaluate the quality of these results, a comparison point is taken with the score obtained with Δ-MFCC features with the same dataset and classifier. Δ-MFCC are computed on short windows of 512 samples, then temporally integrated on larger frames of length 16384 samples with 50 % overlap. Figure 3 clearly indicates that the proposed features reach the same level of precision than the Δ-MFCC for an approximate depth of 5 dB of SRR. It is also clear that there is a sharp error decrease between 1 and 5 dB, and a much slower decay (at least for speech scores) after 5 dB.

Since decompositions at 5 dB are much faster and more sparse than at 10 dB and given that performances do not seem very different, a decomposition depth of 5 dB has been chosen in the remaining of this paper as a good sparsity/representation compromise.

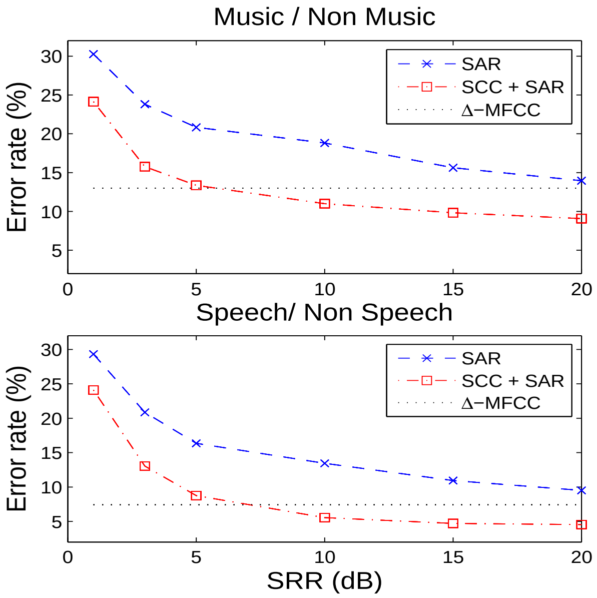

Figure 3: ***Error rates for a Music/non Music and Speech/ non Speech discrimination task over 1 hour of Radio Broadcast and varying SRR. Mean errors obtained with a LDA classifier and random permutations of the Learning and Testing sets***

4.2 Validation on larger datasets

The previous experiment was conducted on just one hour of radio broadcast recordings. It was sufficient to postulate that with a very raw approximation of a signal (5 dB of SRR), robust indexing tasks can be achieved with results comparable to classical methods. To validate this hypothesis, larger and more complicated datasets must be used.

Since finer estimation of the discriminating power is also important, the next experiment will distinguish between music, speech and mixture cases. To do so, the QUAERO evaluation set containing 36 hours of TV and Radio broadcast is used. This set is once again divided in learning and testing parts with a 80 - 20 % ratio and 20 random permutations are used. Table 2 gives the results. They confirm our assumptions that the proposed features reach levels comparable to those of MFCC and Δ-MFCC.

Results emphasize the fact that the proposed features yield better performances than MFCC but poorer ones than Δ-MFCC. It also stresses the fact that they fail to correctly recognise ambiguous situations where speech is present over music. This can easily be explained considering the very low level of decomposition that is used. Mixture cases are often strong speaker voice over faint music or radio jingle and in such cases, the representation at only 5 dB of SRR does not contain enough elements from the music component.

In the last experiment, the whole QUAERO dataset - more than 80 hours of TV and Radio recordings - is used as

		Ground Truth (proportions)		
Features (size)	Classification output	Music (39.53 %)	Speech (38.63 %)	Mix (15.95 %)
MFCC (13)	Music	73.27 (± 4.13 %)	17.49 (± 4.77 %)	33.47 (± 5.89 %)
	Speech	13.80 (± 2.76 %)	67.21 (± 7.16 %)	40.54 (± 4.91 %)
	Mix	6.30 (± 2.11 %)	8.99 (± 2.65 %)	**22.59** (± 3.12 %)
Δ-MFCC (13)	Music	**92.25** (± 1.89 %)	10.75 (± 6.41 %)	19.11 (± 2.95 %)
	Speech	*3.78* (± 1.09 %)	82.47 (± 6.80 %)	65.59 (± 4.10 %)
	Mix	3.12 (± 0.84 %)	5.97 (± 1.70 %)	14.18 (± 1.98 %)
SAR (7)	Music	78.03 (± 5.69 %)	7.77 (± 1.39 %)	19.59 (± 3.21 %)
	Speech	12.30 (± 4.08 %)	81.39 (± 2.27 %)	64.84 (± 4.46 %)
	Mix	6.41 (± 1.33 %)	6.40 (± 0.75 %)	11.93 (± 2.28 %)
SCC + SAR (15)	Music	80.34 (± 5.12 %)	*6.13* (± 1.53 %)	16.17 (± 2.69 %)
	Speech	10.58 (± 3.55 %)	**85.14** (± 2.77 %)	66.44 (± 4.20 %)
	Mix	7.03 (± 1.41 %)	6.63 (± 1.16 %)	15.30 (± 2.40 %)

Table 2: ***Mean scores (and variances) for a speech/music/mixture discrimination task using different features and a LDA classifier. Learning is performed with 20 random permutations of 80% of the dataset. Bold scores indicates the best correct classification scores and italic scores the least critical errors.***

well as another classifier based on Support Vector Machines (SVM). Table 3 gives the speech/music confusion matrix for different features combinations. SAR and SCC features are still computed at 5 dB SRR. Here, performances are slightly worse for the proposed features than for MFCC, but a good overall F-measure of 80 % is reached for music, better than sole Δ-MFCC. Interesting results can be obtained with combinations of new features and MFCC (see for example MFCC + SAR).

Features (Size)		Speech	Music
MFCC (13)	Speech	75.45 %	24.55 %
	Music	14.39 %	85.61 %
Δ-MFCC (13)	Speech	85.10 %	14.90 %
	Music	28.72 %	71.28 %
SCC + SAR (15)	Speech	72.45 %	27.55 %
	Music	17.72 %	80.28 %
MFCC + SAR (20)	Speech	82.45 %	17.55 %
	Music	13.12 %	86.88 %

Table 3: ***Confusion Matrix for different combinations of features with a SVM classifier over the QUAERO project audio database.***

5. CONCLUSIONS AND FUTURE WORK

A compromise between sparsity and speech/music discriminating power has been proposed. The experiments conducted here suggest that, at only 5 dB of Signal-to-Residual Ratio, which is very sparse for most music signals, meaningful features can be derived at very low complexity from the sparse representation, that gives performances comparable to PCM-based methods. Experiments suggest that increasing the decomposition depth yields even better results but this would come at a cost in computational time and storage space. Results also prove that complex tasks can be adressed easily in the sparse domain for large datasets and future work will explore other information retrieval tasks such as similarity search.

Finally, it should be noted that, at low approximation levels such as the 5 dB SRR employed here, the overall sound quality is poor. Perceptual studies should investigate whether a similar tradeoff between sparsity and discriminatory power also holds for human-based source classification.

6. REFERENCES

[1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] S. Kiranyaz, M. Aubazac, and M. Gabbouj. Unsupervized segmentation and classification over MP3 and AAC bitstreams. *Digital Media Processing for multimedia interactive services*, 1:338–344, 2003.

[3] 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*, 16:116, 2008.

[4] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, 41(12):3397–3415, December 1993.

[5] J. Pinquier, J.-L. Rouas, and R. Andre-Obrecht. Robust speech/music classification in audio documents. In *International Conference on Spoken Language Processing*, 2002.

[6] E. Ravelli, G. Richard, and L. Daudet. Union of MDCT bases for audio coding. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1361–1372, 2008.

[7] E. Ravelli, G. Richard, and L. Daudet. Audio signal representations for indexing in the transform domain. *IEEE Trans. on Audio, Speech and Language Processing*, 18:434 – 446, 2010.

[8] J. Saunders. Real-time discrimination of broadcast speech/music. *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, 2:993–996, 1996.

[9] E. Sheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, 2:1331–1334, 1997.

[10] F. Tzanetakis, G. Cook. Sound analysis using MPEG compressed audio. *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, 2:761–764, 2000.