Pattern Recognition Letters 31 (2010) 1489–1497

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Incorporating scale information with cepstral features: Experiments on musical instrument recognition

Marcela Morvidone [a,b], Bob L. Sturm [b], Laurent Daudet [c,*]

[a] Laboratoire de Physique Théorique et Modélisation, Université de Cergy-Pontoise, Site de Saint-Martin, Pontoise, 2 Av. Adolphe Chauvin, 95302 Cergy-Pontoise, France
[b] Institut Jean Le Rond d'Alembert (IJLRA), Équipe Lutheries, Acoustique, Musique (LAM), Université Pierre et Marie Curie (UPMC) – Paris 6, UMR 7910, 11, rue de Lourmel, 75015 Paris, France
[c] Institut Langevin (LOA), Université Paris Diderot – Paris 7, UMR 7587, 10, rue Vauquelin, 75231 Paris, France

ARTICLE INFO

Article history:
Available online 4 January 2010

Keywords:
Audio signal classification
Sparse decompositions
Time–frequency/time-scale features
Musical instrument recognition

ABSTRACT

We present two sets of novel features that combine multiscale representations of signals with the compact timbral description of Mel-frequency cepstral coefficients (MFCCs). We define one set of features, *OverCs*, from overcomplete transforms at multiple scales. We define the second set of features, *SparCs*, from a signal model found by sparse approximation. We compare the descriptiveness of our features against that of MFCCs by performing two simple tasks: pairwise musical instrument discrimination, and musical instrument classification. Our tests show that both OverCs and SparCs improve the characterization of the global timbre and local stationarity of an audio signal than do mean MFCCs with respect to these tasks.

1. Introduction

For speech signal processing tasks, for instance, speaker verification (Bimbot et al., 2004; Ganchev et al., 2005), or speech recognition (Rabiner and Juang, 1993), the use of Mel-frequency cepstral coefficient features (MFCCs) has proven extremely effective because they provide compact and perceptually meaningful descriptions of the distribution of formants. Though the source-filter model does not fit every manner of sound production, the use of MFCCs has also benefited tasks involving discrimination between timbres in non-speech signals, such as musical instrument recognition (Essid et al., 2006; Joder et al., 2009), musical fingerprinting (Casey et al., 2008), and environmental sound classification (Cowling and Sitte, 2003; Couvreur and Laniray, 2004; Defréville et al., 2006). To improve performance in many of these tasks, MFCCs are often combined with other features, such as the set of time- and frequency-domain features specified in the MPEG-7 audio standard (Manjunath et al., 2002).

To describe a signal with time-varying statistics in terms of MFCCs, one computes them in a time-localized fashion over short windows during which the signal is assumed to be stationary, e.g., typically 20–30 ms windows spaced every 10 ms for speech signals. This duration is reasonably based for speech processing on the physics of speech production (Rabiner and Juang, 1993), i.e., the human voice is limited in the number of timbres it can produce per unit time. However, the assumption of stationarity over a single short-time duration for music signals is often unreasonable if we consider the phenomena that occur over a range of different time-scales, for instance, transients, vibrato, tremolo, sustained harmonics, etc. Furthermore, many phenomena can occur simultaneously, which are inseparable in MFCCs by their non-linearity. It does not make sense to think that the MFCCs of a portion of a music signal that contains a strong transient are meaningful other than to say the power spectral density is more wideband than in another segment. In such a case, it would be more perceptually relevant to separate the description of the transient (Herrera-Boyer et al., 2003) from that of the rest of the local signal.

For recognition tasks involving signals with phenomena occurring over multiple time-scales, such as musical signals (Cowling and Sitte, 2003; Couvreur and Laniray, 2004; Defréville et al., 2006; Aucouturier et al., 2007; Casey et al., 2008; Joder et al., 2009), MFCCs generated using a single window size and uniform translations must be suboptimal with respect to their descriptiveness, than are features computed with considerations for the diversity of time-scales present. What is desired for such signals then is a set of features that have the compact descriptiveness of MFCCs while taking into account the local statistics of the signal so that one may better distinguish its contents. In other words, for signal content having slowly varying statistics it is unnecessary to compute MFCCs every 10 ms using a 30 ms window; and for signal content occurring over short time-scales we should avoid "smearing" its influence over a large window. We want to separate the

* Corresponding author. Tel.: +33 (0) 1 40 79 46 92; fax: +33 (0) 1 40 79 44 68.
E-mail address: laurent.daudet@espci.fr (L. Daudet).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.12.035

influence of these contents so that the signal features are more local and specific, and can better characterize a signal.

In this paper we address these issues by defining and testing two sets of novel features that provide information about the spectral shape of a signal, as well as the time-scales involved. One set of features, *OverCs* (pronounced "over seas"), basically aggregates the MFCCs computed from redundant transforms performed at multiple scales. The other set of features, *SparCs* (pronounced "spar seas"), looks at the distribution of energy among frequency and scale of a sparse model of a signal found using sparse approximation and a multiscale time–frequency dictionary. Sparse approximation methods have been shown to provide efficient and meaningful parametric representations adapted to audio data (Mallat and Zhang, 1993; Gribonval and Bacry, 2003; Daudet, 2006; Leveau et al., 2008), where small-scale atoms are used to model transients, and large-scale atoms are used for tonals. This provides a level of source separation that can remove the influence of short-scale phenomena on the spectral description of large-scale phenomena. Though sparse approximation has a large computational complexity, we consider the case where a database of audio signals has already been decomposed, for instance, in its compression (Ravelli et al., 2008). We expect that both of these features will be more discriminative than mean MFCCs for content recognition tasks because they consider multiple scales. We test this by comparing the discriminative ability of these features with those of mean MFCCs in two simple tasks: pairwise musical instrument discrimination, and musical instrument classification. For each of these features, we select a small number of coefficients to compare with the mean MFCCs at a fixed dimensionality. The signals we use are excerpted from real musical recordings, and have been used in other automatic classification works (Essid, 2005). We find that for both tasks our proposed features perform better than the mean MFCCs.

The rest of this paper is organized as follows. In Section 2 we review MFCCs, and then sparse approximation with time–frequency dictionaries. We define our new features in Section 3, and provide some examples. In Section 4 we detail our experiments and discuss the results and their significance. Finally, we conclude with a description of ongoing work in several directions.

2. Background

In this section, we review MFCCs, as well as sparse approximation using greedy iterative methods with a time–frequency dictionary.

2.1. Review of Mel-frequency Cepstral Coefficients (MFCCs)

The *cepstrum* models the distribution of spectral energy in a time-domain signal. Given a real length-N discrete signal, x, defined over $0 \leqslant n < N$, and its discrete Fourier transform (DFT), $X = DFT\{x\}$, its *real cepstrum* is defined (Rabiner and Juang, 1993)

$$c[l] \triangleq DFT^{-1}\{\log|X|\} = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}\log|X[k]|e^{j2\pi kl/N}, \quad l = 0, 1, \ldots, N-1 \tag{1}$$

where the logarithm is used to separate the filtering from the excitation signal. The magnitudes of the cepstrum provide a description of the spectral shape of x independent of a wideband excitation signal. Because this fits the manner of speech production, the cepstrum has been very successful for tasks of speaker and speech recognition (Rabiner and Juang, 1993).

To create a less redundant and perceptually-relevant representation of the spectral shape of x, one finds the energies in frequency bands according to perceived pitch by the *Mel-frequency scaling*, which maps a frequency f (Hz) to Mels by $\phi(f) = 1127\ln(1 + f/700)$. The filterbank we use has $L = 40$ overlapping bands with triangular magnitude responses, weighted such that each has equal area, beginning at a low frequency of 133.33 Hz, and ending at the high frequency of 6853.84 Hz (Ganchev et al., 2005). The first 13 filters are linearly spaced every 66.67 Hz with a bandwidth of 133.33 Hz, and weighting 0.015. The last 27 filters have center frequencies that are logarithmically spaced from 1073.4 Hz to 6413.59 Hz. The center frequencies of the filters are given by

$$f_c(l) = \begin{cases} 133.33 + 66.66l, & l = 1, 2, \ldots, 13 \\ 1073.4(1.0711703)^{(l-14)}, & l = 14, 15, \ldots, 40. \end{cases} \tag{2}$$

Each filter ($l = 1, 2, \ldots, 40$) is given by

$$H_l[k] \triangleq \begin{cases} 0, & 0 \leqslant kF_s/N < f_c(l-1) \\ \frac{kF_s/N - f_c(l-1)}{f_c(l) - f_c(l-1)}, & f_c(l-1) \leqslant kF_s/N < f_c(l) \\ \frac{kF_s/N - f_c(l+1)}{f_c(l) - f_c(l+1)}, & f_c(l) \leqslant kF_s/N < f_c(l+1) \\ 0, & f_c(l+1) \leqslant kF_s/NF_s \end{cases}, \tag{3}$$

where F_s is the Nyquist sampling rate (Hz), $f_c(0) \triangleq 133.33, f_c(41) \triangleq 6853.84$, and the band-dependent magnitude factors $\{a_l\}$ are given by

$$a_l \triangleq \begin{cases} 0.015, & 1 \leqslant l \leqslant 13 \\ \frac{2}{f_c(l+1) - f_c(l-1)}, & 14 \leqslant l \leqslant 40. \end{cases} \tag{4}$$

With this filter bank, we compute the *Mel-frequency cepstral coefficients* (MFCCs) by a discrete cosine transform (DCT)

$$cc_M[m] \triangleq \beta_L(m)\sum_{l=1}^{L}\left(\sum_{k=0}^{P-1}\log|X[k]a_lH_l[k]|\right)\cos\left[\frac{m\pi}{L}\left(l - \frac{1}{2}\right)\right], \tag{5}$$

defined for $0 \leqslant m < L$, and where the normalization factor is defined

$$\beta_R(y) \triangleq \begin{cases} 1/\sqrt{R}, & y = 0 \\ \sqrt{2/R}, & 1 \leqslant y \leqslant R. \end{cases} \tag{6}$$

For speech and audio data, the DCT provides a satisfactory decoupling of the components of their log magnitude spectra (Logan, 2000). Since speech and audio signals are non-stationary, MFCCs are calculated in practice using overlapping sliding windows. We define the short-time MFCCs

$$cc_M[m, p] \triangleq \beta_L(m)\sum_{l=1}^{L}\left(\sum_{k=0}^{N-1}\log|X[k, p]a_lH_l[k]|\right)\cos\left[\frac{m\pi}{L}\left(l - \frac{1}{2}\right)\right], \tag{7}$$

for $0 \leqslant m < L$, and where the DFT of x localized at time p is defined

$$X[k, p] \triangleq \frac{1}{\sqrt{P}}\sum_{n=0}^{P-1}x[n + p]w[n]e^{-j2\pi kn/P}, \quad 0 \leqslant k \leqslant P - 1, \tag{8}$$

for time shifts $0 \leqslant p < N - P$, and a real window w that is non-zero for $0 \leqslant n < P$, and that satisfies

$$\frac{1}{\sqrt{P}}\sum_{n=0}^{P-1}|w[n]|^2 = 1. \tag{9}$$

In speech recognition and music processing it is typical to use windows of length 20 – 30 ms, over which durations the signal can be said to be approximately stationary. These windows are usually placed at translations of half their duration. For speech signals, only the first $M = 13$ coefficients are typically kept (Rabiner and Juang, 1993), excepting the term at $m = 0$ since it reflects only the short-term energy of the signal.

2.2. Sparse Approximation with Time–Frequency Dictionaries

Sparse approximation attempts to find a small subset of M atoms in a *dictionary* $\mathscr{D} = \{g_\gamma\}_{\gamma\in\Gamma}$ to approximate the length-N function x to some specified maximum error $\epsilon \geqslant 0$, e.g.,

$$\min M \text{ subject to } \left\| x - \sum_{m=1}^{M} \alpha_{\gamma_m} g_{\gamma_m} \right\|_2^2 \leqslant \epsilon, \tag{10}$$

where $\{\gamma_m\}_{m=1}^M \subset \Gamma$, and $\{\alpha_{\gamma_m}\}_{m=1}^M$ is the set of weights associated with the M atoms selected in $\mathscr{D}$. The term *sparse* refers to the desirable property by which the number of atoms selected $M \ll N$, the dimension of the signal. Usually, $|\Gamma| \gg N$ such that many possible solutions exist.

Various methods have been proposed to solve (10), e.g., a greedy iterative strategy (Mallat and Zhang, 1993), or convex optimization with a relaxed sparsity constraint (Chen et al., 1998). The method we use in this paper is Matching Pursuit (MP) (Mallat and Zhang, 1993). MP selects the $(M+1)$th atom from $\mathscr{D}$ by the criterion

$$\gamma_{M+1} = \arg\min_{\gamma\in\Gamma} \left\| R^M x - \frac{\langle R^M x, g_\gamma\rangle g_\gamma}{\|g_\gamma\|_2^2} \right\|_2^2 = \arg\max_{\gamma\in\Gamma} \frac{|\langle R^M x, g_\gamma\rangle|}{\|g_\gamma\|_2}, \tag{11}$$

where the inner product between two real length-N functions x and y is defined

$$\langle x, y\rangle \triangleq \sum_{n=0}^{N-1} x[n]y[n], \tag{12}$$

and the Mth-order residual signal is defined

$$R^M x \triangleq x - \sum_{m=1}^{M} \alpha_{\gamma_m} g_{\gamma_m}. \tag{13}$$

MP defines the $(M+1)$th weight by

$$\alpha_{\gamma_{M+1}} \triangleq \frac{\langle R^M x, g_{\gamma_{M+1}}\rangle}{\|g_{\gamma_{M+1}}\|_2^2}. \tag{14}$$

In this way, MP iteratively selects the atom that maximally reduces the ℓ_2-norm of an intermediate residual, beginning with $R^0 x = x$.

Various dictionaries have been used in the sparse approximation of speech and audio signals, such as Gabor time–frequency atoms (Mallat and Zhang, 1993), harmonic atoms (Gribonval and Bacry, 2003; Leveau et al., 2008), or unions of bases, such as cosine and wavelet bases (Daudet, 2006), and multiscale cosine bases (Ravelli et al., 2008). In this work, we use time–frequency atoms that have a single modulation frequency ω and phase ϕ, time scale s, and time shift u. Considering that an element $\gamma \in \Gamma$ is the quadruple $\gamma_m = (s_m,\ u_m,\ \omega_m,\ \phi_m)$, a real length-$N$ time–frequency atom is defined

$$g_{\gamma_m}[n] \triangleq g\left(\frac{n-u_m}{s_m}\right)\cos(\omega_m n + \phi_m), \quad 0 \leqslant n < N, \tag{15}$$

where $g(t)$ is a continuous prototype lowpass function, for instance a Gaussian window. The dictionary specified by Table 1 contains atoms of scale 128 samples located at translations that are integer multiples of 32 samples i.e., $u_m = 32l,\ l \in \mathbb{Z}$, and with modulation frequencies that are positive integer multiples of 43.1 Hz, i.e., $\omega_m = 2\pi l 43.1/F_s,\ l \in \mathbb{Z}^+$ (up to the Nyquist frequency πF_s). When the dictionary $\mathscr{D}$ is complete, then $\lim_{M\to\infty}\|R^M x\|_2^2 = 0$, i.e., the representation converges such that the error becomes zero (Mallat and Zhang, 1993). In this article, we use MP Toolkit (Krstulovic and Gribonval, 2006) and Gaussian lowpass functions.

Table 1
Dictionary parameters of a scaled, shifted, and modulated Gaussian lowpass function: scale s, time resolution Δ_u, and frequency resolution Δ_f. Durations are specific to a sampling rate $F_s = 44100$ Hz.

s (samples/ms)	Δ_u (samples/ms)	Δ_f (Hz)
128/2.9	32/0.7	43.1
256/5.8	64/1.5	43.1
512/11.6	128/2.9	43.1
1024/23.2	256/5.8	43.1
2048/46.4	512/11.6	21.5
4096/92.9	1024/23.2	10.8
8192/185.8	2048/46.4	5.4
16384/371.5	4096/92.9	2.7

3. Incorporating Scale Information with Cepstral Features

We now define two sets of novel features, OverCs and SparCs, that combine a time-domain scale characterization of a signal with a frequency-domain spectral characterization. For OverCs, we use overcomplete transforms performed at multiple scales. For SparCs, we use a signal model found by sparse approximation and a multiscale time–frequency dictionary. Sparse approximation allows a minimum amount of source separation between phenomena that occur over different time scales. Though OverCs are less computationally intensive than SparCs, it does not allow for any source separation.

3.1. OverCs: MFCC-like Features from Overcomplete Transforms

We generate OverCs by first computing short-time MFCCs (7) using windows of multiple scales. The parameters of the windows we use are the same as for the overcomplete dictionary detailed in Table 1. This means that we essentially project x onto all atoms of this dictionary, and then use these values to compute mean MFCCs for each scale. Let us define $cc_M[m,p,s]$ to be (7) computed using a window scale $128\cdot 2^{(s-1)}$, which are the L MFCCs of x over the time region $[p, p+128\cdot 2^{(s-1)})$. Now we define the set of short-time MFCCs

$$\mathscr{C}_{s,\epsilon} \triangleq \{cc_M[m,p,s] : cc_M[0,p,s] > \epsilon, \quad 1 \leqslant m < 40\}, \tag{16}$$

from time regions of scale index s in x with energy greater than $\epsilon \geqslant 0$ (to avoid signal frames near to silence), and where we have excluded all $m = 0$ cepstral coefficients since they only contain energy information. We then form averages of this set

$$\overline{cc_M}[m,s] \triangleq \frac{1}{|\mathscr{C}_{s,\epsilon}|}\sum_{p\in\mathscr{C}_{s,\epsilon}} cc_M[m,p,s], \tag{17}$$

where $p \in \mathscr{C}_{s,\epsilon}$ means those p that exist in $\mathscr{C}_{s,\epsilon}$. In other words, we average all contributions in cepstral index m and scale index s over the short-time frames of size s. Finally, since there will be redundancy across scales in each cepstral index, we perform a DCT in the scale direction to generate the OverCs:

$$\zeta[m,z] \triangleq \beta_8(z)\sum_{s=1}^{8} \overline{cc_M}[m,s]\cos\left[\frac{(z-1)\pi}{8}\left(s-\frac{1}{2}\right)\right], \tag{18}$$

for $1 \leqslant z \leqslant 8$, and using the normalization factor (6).

3.2. SparCs: MFCC-like Features from Sparse Models

Consider that we have approximated x by M atoms selected from $\mathscr{D}$ by MP, thus producing an M-order model (10). We first separate the M atoms in the sparse model based on their modulation frequency and scale parameters. Let us associate each atom with an integer and define this index set $\mathscr{M} \triangleq \{1, 2, \ldots, M\}$. Thus, atom $m \in \mathscr{M}$ has parameters $(s_m,\ u_m,\ \omega_m,\ \phi_m)$. We define the following

mappings from the scale-frequency space of the model. For the dictionary specified in Table 1, we map each atom scale to an integer by

$$S(s) \triangleq 1 + \log_2(s/128), s \in \{128, 256, 512, 1024, 2048, 4096, 8192, 16384\}. \tag{19}$$

Based on the 40-band filterbank described in Section 2.1, we define the modulation frequency mapping

$$W(\omega) \triangleq \begin{cases} l, & f_c(l-1) \leqslant \Omega\omega < f_c(l+1), \quad 1 \leqslant l \leqslant 39 \\ 40, & f_c(40) \leqslant \Omega\omega < f_c(41) \end{cases}, \tag{20}$$

where $\Omega \triangleq F_s/2\pi$, and the center frequencies are given by (2). Now, we define the index set of atoms in the sparse model that have some scale index σ and modulation frequency index l as

$$\mathscr{M}_{l\sigma} \triangleq \{m \in \mathscr{M} : W(\omega_m) = l, \quad S(s_m) = \sigma\}. \tag{21}$$

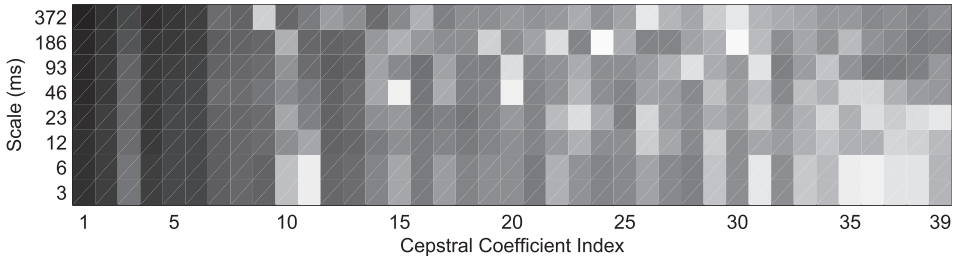
Note that there is no longer any notion of where in time a given atom exists.

From the set $\{\mathscr{M}_{l\sigma} : 1 \leqslant l \leqslant 40,\ 1 \leqslant \sigma \leqslant 8\}$, we accumulate the magnitude weights as a function of modulation frequency and scale:

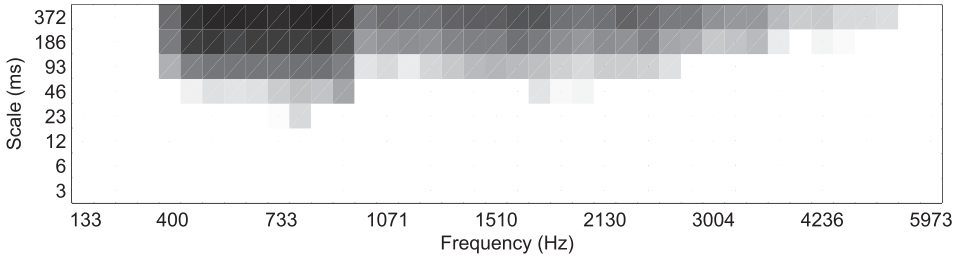
$$X[l, \sigma] \triangleq \sum_{m \in \mathscr{M}_{l\sigma}} a_l H_l(\omega_m) |\alpha_{\gamma_m}|, \tag{22}$$

where the filter weights $\{a_l\}$ are given by (4), and we define, similar to (3),

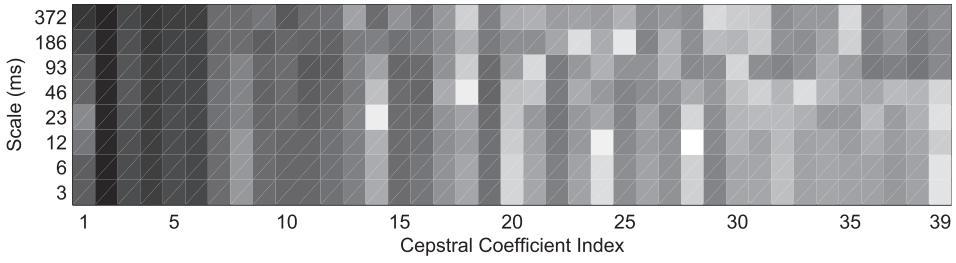
(a) $|\overline{cc_M}(m, s)|$ (17) for Clarinet



(b) $|\log X(\sigma, l)|$ (22) for Clarinet



(c) $|\overline{cc_M}(m, s)|$ (17) for Trumpet



(d) $|\log X(\sigma, l)|$ (22) for Trumpet

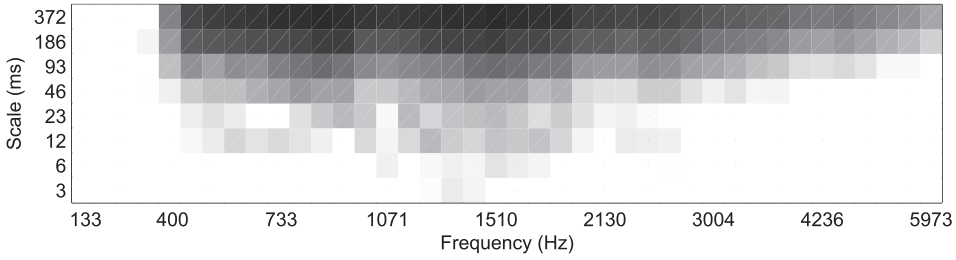


Fig. 1. $\overline{cc_M}[m, s]$ (17) and $|\log X(\sigma, l)|$ (22) for two musical instruments playing a chromatic scale from C5 to B5.

$$H_l(\omega) \triangleq \begin{cases} \frac{\Omega\omega - f_c(l-1)}{f_c(l) - f_c(l-1)}, & f_c(l-1) \leqslant \Omega\omega < f_c(l) \\ \frac{\Omega\omega - f_c(l+1)}{f_c(l) - f_c(l+1)}, & f_c(l) \leqslant \Omega\omega < f_c(l+1), \\ 0, & else \end{cases} \tag{23}$$

to weight the contribution of the atom to the lth frequency bin of the relevant scale index. We take the two-dimensional DCT of $X[l,\sigma]$, which, for this dictionary of 8 scales and filterbank of 40 bands, is defined

$$\xi[m,z] \triangleq \sum_{l=1}^{40}\sum_{\sigma=1}^{8} X[l,\sigma] \times \beta_{40}(m) \times \cos\left[\frac{(m-1)\pi}{40}\left(l-\frac{1}{2}\right)\right]\beta_8(z)\cos\left[\frac{(z-1)\pi}{8}\left(\sigma-\frac{1}{2}\right)\right], \tag{24}$$

for $1 \leqslant m \leqslant 40$ and $1 \leqslant z \leqslant 8$, with the normalization factors defined by (6). Finally, we create the SparCs by setting $\xi[1,1] = 0$, and normalizing the rest:

$$\hat{\xi}[m,z] \triangleq \frac{\xi[m,z]}{\sqrt{\sum_{m=1}^{40}\sum_{z=1}^{8}|\xi[m,z]|^2}}. \tag{25}$$

SparCs are very fast to compute if a sparse decomposition is already available.

3.3. Examples of SparCs and OverCs

Fig. 1(a) and (c) show the magnitude mean short-time MFCCs (17) for two different musical audio signals recorded from a clarinet and trumpet (IOWA, 2009). Each instrument plays an

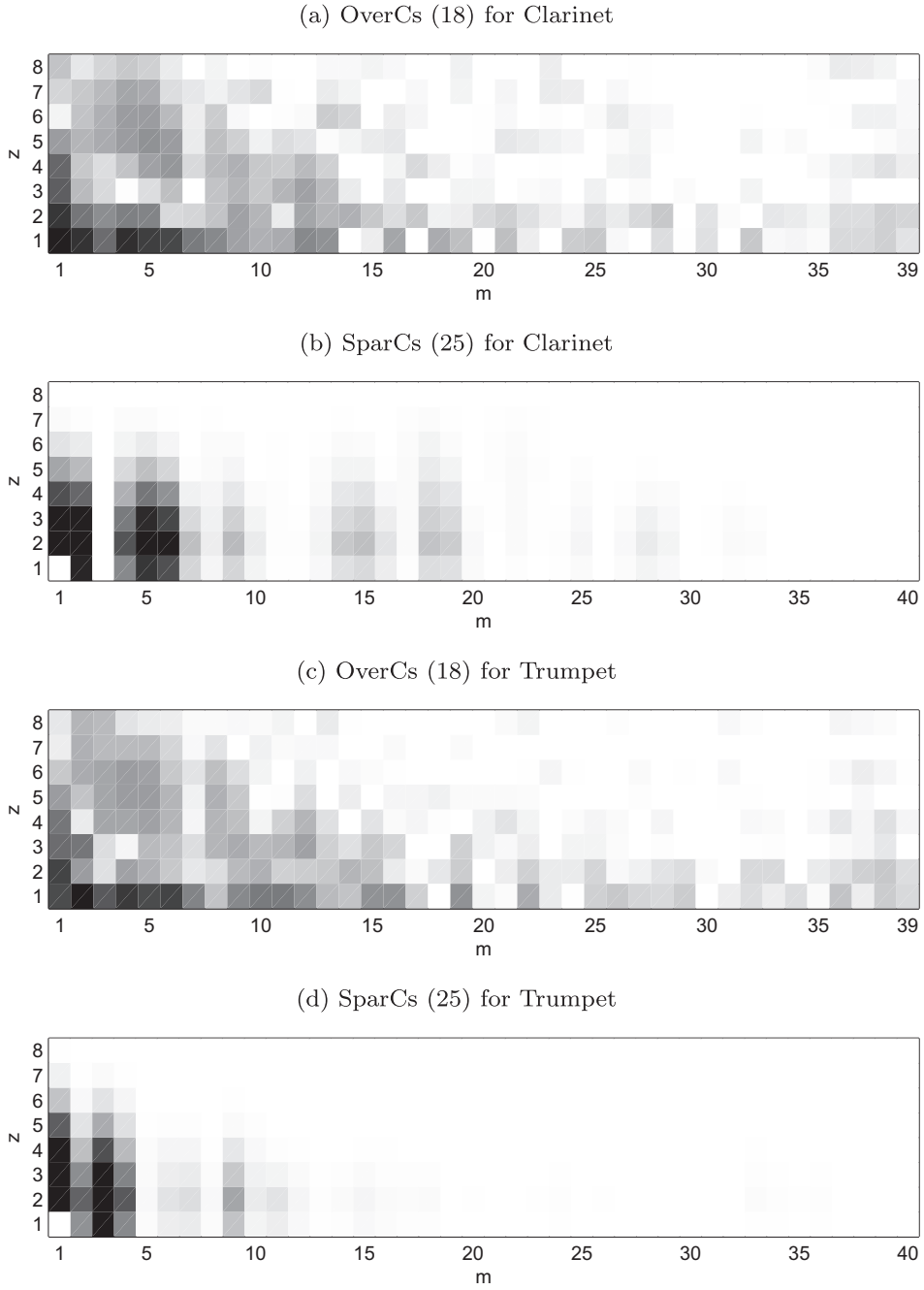


Fig. 2. OverCs $|\zeta[m,z]|$ (18) and SparCs $|\hat{\xi}[m,z]|$ (25) from data in Fig. 1. Mean MFCCs of these two signals are shown in Fig. 3.

ascending chromatic scale (C5 to B5), lasting 32 s (clarinet) or 97 s (trumpet). Notice that each row contains the average short-time MFCCs over the entire signal calculated with a window of one scale. Each column displays how the average short-time MFCCs change as a function of scale over the entire signal at a particular cepstral index m. For the same signals we see in Fig. 1(b) and (d) the distribution of energy among atom scale and frequency, $X[l,\sigma]$ (22) from their sparse approximations. Of the two, the trumpet clearly has a wider bandwidth in frequency and scale. Note that MP does not perform any windowing in the decomposition.

For these same examples, Fig. 2(a) and (c) show their OverCs (18), and Fig. 2(b) and (d) show their SparCs (25). We see that the SparCs are much more compact than the OverCs. The elements $\zeta[m,1]$ and $\hat{\xi}[m,1]$ are most closely relatable to mean MFCCs since they are an average of the cepstral coefficients observed over all eight scales. We show these for both instruments in Fig. 3. The fact that large coefficients are concentrated in different regions of the transformed frequency-scale space motivates the idea that SparCs and OverCs can be useful for classifying sound sources.

In summary, the OverCs of a signal are obtained by computing short-time MFCCs (7) with multiple window sizes — in essence projecting the signal onto all elements of a multiscale time–frequency dictionary and then filtering and computing the DCT for each scale — and then decoupling the MFCCs over scale by performing a DCT in the scale direction (18). The SparCs of a signal are obtained by first performing a sparse approximation over a multiscale time–frequency dictionary, then building the frequency-scale function (22) from the parameters of the decomposition, taking the 2-dimensional DCT of this function (24), and then removing the energy term and normalizing the rest (25).

4. Application to Automatic Musical Instrument Recognition

We now discuss our evaluation of the performance of these new sets of features with respect to two tasks of musical instrument recognition. Our musical instrument data consists of a subset of the one described in (Essid, 2005) — containing monophonic data recorded at 44.1 kHz, some of which belong to the RWC database (Goto et al., 2002), and some recordings made by Essid et al. (Essid, 2005) — and some data extracted ourselves from commercial CDs. Our database consists of seven instruments, that represent our *classes*: clarinet (Cl), oboe (Ob), violin (Vl), cello (Co), guitar (Gt), piano (Pn), and trumpet (Tr). For each of these instrument classes we have several five-second excerpts from solo recordings from five different *sources* per class, for instance, different instruments, performers, recording conditions, and music, for a total of 2,755 excerpts. No excerpt from two different sources comes from the same CD. It should be emphasized that these data are extracted from real music performances, and thus are not isolated note recordings. For example, there exist double and triple stops in the violin and cello examples, chords in the guitar and piano examples, pitch bends in the clarinet, as well as reverberation. We summarize our database in Table 2.

We create the SparCs feature database using the dictionary defined in Table 1 with a Gaussian window, to the order M where the signal-to-residual energy ratio reaches $20\log_{10}(\|x\|_2/\|R^M x\|_2) = 30$ dB. We create the OverCs feature database with the same parameters as in Table 1. From these features, we select three different subsets of 13 (m,z) features:

$$\mathscr{I}_1 \triangleq \{(2,1),\ldots(14,1)\} \tag{26}$$

$$\mathscr{I}_2 \triangleq \{(2,1),\ldots(9,1),\ (1,2),\ \ldots,\ (3,2),\ (1,3),\ (2,3)\} \tag{27}$$

$$\mathscr{I}_3 \triangleq \{(2,1),\ldots(8,1),(1,2),\ \ldots,(3,2),\ (1,3),\ (2,3),\ (1,4)\}, \tag{28}$$

where, for instance, $\text{SparCs}(\mathscr{I}_1) = \{\hat{\xi}[2,1],\ldots\hat{\xi}[14,1]\}$. For the sake of comparison, we also compute and select the first 13 elements (excepting the energy term) of the mean MFCCs feature vector built with Hamming windows of 30 ms scale and 10 ms translations, and keeping only those MFCCs from signal segments where the normalized signal energy exceeds the threshold $\epsilon \geqslant 0.1$. This number of coefficients is a common choice in speech processing (Rabiner and Juang, 1993).

The reasoning behind these feature choices is as follows. Features in $\mathscr{I}_1$ do not consider change in energy as a function of atom scale, and should closely approximate the mean MFCCs, as seen in Fig. 3. We thus expect the performance of this feature set to be close to that of the mean MFCCs. Features in $\mathscr{I}_2$ include a few elements describing how energy is distributed as a function of scale; and those in $\mathscr{I}_3$ include a few more. We expect that including information about energy variation over atom scales will help discriminate instruments that have similar spectral shapes, but different excitations and note transitions. There should be a trade-off in the usefulness of the number of terms selected from each part of the SparCs and OverCs, and certainly some of their components will be more useful than others for discriminating between certain

Table 2
Summary of instrument database, showing the number of 5-s excerpts for each of the five sources, and the total number of excerpts for each instrument class.

Instrument	Label	# sources	# excerpts	Total #
Clarinet	Cl	5	56	280
Cello	Co	5	116	580
Guitar	Gt	5	74	370
Oboe	Ob	5	90	450
Piano	Pn	5	77	385
Trumpet	Tr	5	49	245
Violin	Vl	5	89	445
Total number of examples				2755

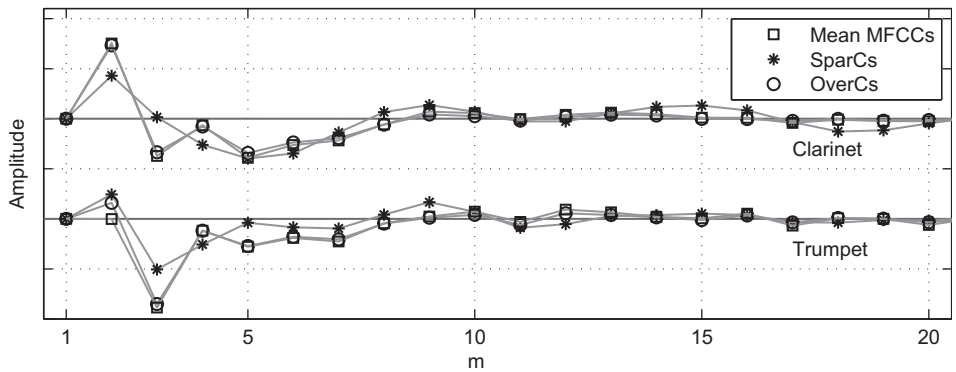


Fig. 3. The first twenty elements of Mean MFCCs, OverCs $\zeta[m,z]$, and SparCs $\hat{\xi}[m,z]$ for $z = 1$, for the data shown in Fig. 2.

groups of instruments. As a first step though, we use the three feature subsets (26)–(28).

To compare the discriminatory ability of each set of features, we use a Support Vector Machine (SVM) with a radial basis function kernel, which was also used in (Essid, 2005) — although there with a much larger set of features, coupled with different feature selection schemes. We perform two different types of tests: pairwise discrimination, and instrument classification. In the pairwise discrimination task we assume the unknown instrument is one of two, and we must determine which one it is. Assume we are discriminating between Ob and Vl. To create the test data we select 49 realizations randomly for one source of Ob, and 49 realizations randomly from one source of Vl, as shown in Fig. 4. We create the training data by choosing 49 realizations randomly from each of the remaining sources of both instruments. In the training stage, two parameters must be fixed: the penalty parameter of the error term C, and a kernel parameter γ. These parameters are optimized by a "grid search" procedure (Hsu et al., 2009), which tests several pairs of (C,γ) and selects the one giving the best 5-fold cross-validation accuracy. The grid is defined with $C \in \{2^i : i = -5,\ -3,\ \ldots,\ 15\}$, $\gamma \in \{2^i : i = -15,\ -13,\ \ldots, 3\}$. Finally, we test the optimized SVM using the testing data. For each instrument class pair, we repeat this procedure ten times for each of the 25 different possible source pairings, and then average the results. The SVM parameters (C,γ) can change for each pair of sources; nevertheless, preliminary tests show that classifier performance is robust to these changes of (C,γ).

In the more general instrument classification task of one-vs-all, we assume the unknown instrument is one of the seven in Table 2. To create the test data in each instrument classification task, we select 49 realizations randomly from one source for the instrument class being tested. To create the training data we select 49 realizations randomly from each of the remaining sources of the instrument being tested, and 49 realizations randomly from each of all five sources of the six other instrument classes. No realizations from the same source appear in both the training and testing data. We train the SVM by the grid search strategy described above. After running the classification task described, we repeat the same procedure ten times for each source of the instrument class we are detecting.

Fig. 5 shows the results of the pairwise instrument discrimination task with respect to the performance of mean MFCCs, for each of the features OverCs($\mathscr{I}_2$) and SparCs($\mathscr{I}_2$). We see that in only two pairings (CoOb, CoTr) does the inclusion of scale information not help the discrimination with respect to using mean MFCCs. For the pairwise instrument discrimination task, we summarize the

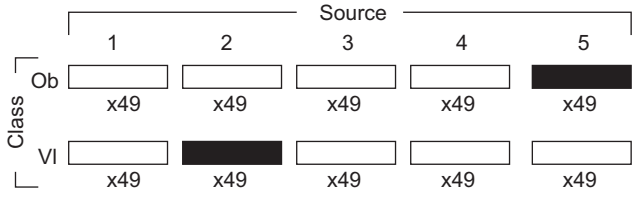


Fig. 4. Example of testing and training data selection in pairwise discrimination task for Ob-Vl. Test data (49 realizations) is selected randomly from sources in black, for each instrument class. Training data (49 realizations) is selected randomly from each of the four remaining sources of each instrument.

Table 4
Confusion matrix for instrument classification using mean MFCCs (top), OverCs($\mathscr{I}_2$) (middle), and SparCs($\mathscr{I}_2$) (bottom) features. Best scores are in bold.

	Cl	Co	Gt	Ob	Pn	Tr	Vl
MFCCs							
Clarinet	72.53	2.45	5.47	5.59	1.59	9.31	3.06
Cello	3.18	70.33	6.94	0.24	4.90	0.041	14.37
Guitar	13.55	3.51	75.18	0	7.63	0	0.12
Oboe	9.43	0.12	0.12	78.16	0	11.96	0.20
Piano	3.51	2.53	9.10	0	84.04	0.82	0
Trumpet	9.59	0	0	12.73	1.43	73.35	2.90
Violin	6.20	12.28	0.61	0.45	0.04	5.35	75.06
OverCs($\mathscr{I}_2$)							
Clarinet	**83.92**	1.18	1.22	4.73	0.45	7.59	0.90
Cello	1.14	**78.90**	4.20	0.61	1.14	0.24	13.75
Guitar	3.67	5.18	**81.06**	0.081	8.69	0	1.30
Oboe	5.88	0.33	0.29	81.02	0	11.88	0.61
Piano	0.49	2.82	9.31	0	**86.69**	0.65	0.04
Trumpet	8.08	0	0	11.92	0.37	**77.10**	2.53
Violin	1.47	13.80	0.73	0.12	0	1.55	**82.33**
SparCs($\mathscr{I}_2$)							
Clarinet	82.98	3.67	1.39	2.33	0.20	7.51	1.92
Cello	6.45	76.61	4.45	0.77	2.12	1.02	8.57
Guitar	0.20	7.59	75.22	0	16.94	0	0.04
Oboe	5.22	0.16	0	**81.18**	0	13.43	0
Piano	3.31	3.35	21.96	0	69.92	0.04	1.43
Trumpet	6.86	0.86	0	10.94	0	76.77	4.57
Violin	5.31	17.67	0.73	0.37	0.41	2.53	72.98

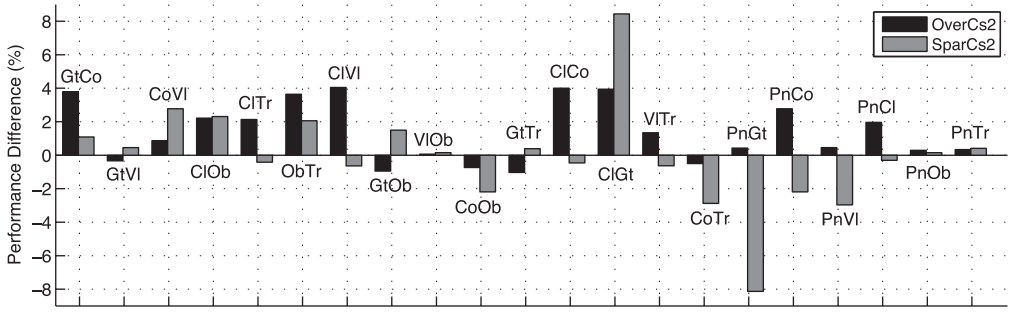


Fig. 5. Correct discrimination rates for all instrument pairs relative to that of the mean MFCCs for OverCs($\mathscr{I}_2$) and SparCs($\mathscr{I}_2$).

Table 3
Correct instrument discrimination rates for each set of features.

	MFCCs	OverCs($\mathscr{I}_1$)	OverCs($\mathscr{I}_2$)	OverCs($\mathscr{I}_3$)	SparCs($\mathscr{I}_1$)	SparCs($\mathscr{I}_2$)	SparCs($\mathscr{I}_3$)
Mean	94.00	94.35	95.37	95.39	91.98	93.95	93.95
Stan. dev.	5.07	4.41	4.23	4.82	6.47	5.22	5.36

Table 5
Mean correct classification rates for all features in instrument classification task.

	MFCCs	OverCs($\mathscr{I}_1$)	OverCs($\mathscr{I}_2$)	OverCs($\mathscr{I}_3$)	SparCs($\mathscr{I}_1$)	SparCs($\mathscr{I}_2$)	SparCs($\mathscr{I}_3$)
Mean	75.52	76.06	81.57	80.52	72.16	76.52	76.76
Stan. dev.	4.48	4.22	3.16	4.98	7.47	4.50	4.61

statistics of the rates of correct discrimination for each of the features in Table 3. We tested the statistical significance of the differences of these means with an ANOVA test. We find no significant difference in the performance of mean MFCCs and SparCs($\mathscr{I}_2$)($p \approx 0.3$). On the other hand, OverCs($\mathscr{I}_2$) perform better than MFCCs on average ($p < 10^{-16}$).

We summarize the results of the instrument classification task in Table 4 as confusion matrices, i.e., percentages of each classification using the mean MFCCs, OverCs($\mathscr{I}_2$) and SparCs($\mathscr{I}_2$) features. For example, in the first line we see that the SVM trained to recognize clarinet using the mean MFCCs feature correctly classified 72.53% of the clarinet realizations presented, but classified 2.45% of them as cello. From these tables, we see that we obtain the best classification rates here with the OverCs($\mathscr{I}_2$) features.

In Table 5 we show the average rate of correct classification for every feature. ANOVA tests show that SparCs($\mathscr{I}_2$) perform marginally better than mean MFCCs on average with $p \approx 0.03$ but they perform significantly better for some instruments (Cl, Co with $p < 10^{-6}$). OverCs($\mathscr{I}_2$) features perform systematically better than mean MFCCs features. For most of the instruments this difference is highly significant ($p < 10^{-6}$), whereas it is less so for Tr ($p \approx 10^{-3}$) and for Ob, Pn ($p \approx 10^{-2}$). For Cl, Co, Gt and Vl, adding the scale information significantly improves the recognition task by 6–11% with respect to mean MFCCs ($p < 10^{-6}$). Of all instruments, we find the least gain in classification for Pn — only 2.65% ($p \approx 0.0086$). It should be noted that the piano is the only percussive instrument in our database. Its spectrum has a regular, decaying behavior except in its attacks. Thus, adding scale information is probably irrelevant in this case. For instruments with sustained sound, however, the scale parameters could be capturing some of the fine details of the timbre, for instance, loudness stability, and playing techniques, for instance, vibrato.

5. Conclusion and Future Work

We have presented two sets of new features, OverCs and SparCs, that combine the compact timbral descriptiveness of MFCCs with multiscale representations of signals. OverCs are created from considering the mean MFCCs calculated over multiple time-scales. SparCs are created from sparse representations created with Matching Pursuit (MP) and multiscale time–frequency dictionaries. With these features we seek to overcome some of the inherent limitations of MFCCs computed using a single window size for non-stationary signals that have a variety of content occurring over different time-scales, such as audio and music signals.

We tested these new features in two simple tasks: pairwise musical instrument discrimination, and musical instrument classification. The results of these tests show that our features outperform mean MFCCs, and in some cases by a significant amount. By selecting only 13 coefficients from these features, classification rates are close to state-of-the-art automatic instrument identification, where best results are usually obtained with highly optimized classifiers and high-dimensional feature selection steps. For example, (Essid, 2005) reports typical correct classification rates of 80–90% using 160 features per audio segment and an optimized SVM kernel. (Note that this different experimental setup and database make comparisons with the work in this paper difficult.) While OverCs appear to outperform SparCs in most cases, SparCs can be computed very quickly when the sparse representations have already been found.

Our current research is examining a number of issues raised by this study. First, what is the optimal choice of features for classification purposes? It is likely that some performance could be gained by using feature selection techniques to determine the trade-off between the number of features (across the frequency and scale indices) and the performance and robustness of classifiers. Also, for the sake of comparison with standard MFCC methods we have only kept 13 coefficients in our feature vectors; further experimentation will help determine the best subset of coefficients from our new features for a given task.

Second, are there other types of signals where these scale-dependent features are useful? Our preliminary experiments in speaker recognition and spoken digit recognition using SparCs indicate that there is not much to be gained by using scale-dependent models over even simple vector quantization strategies using MFCCs. While the structure of speech signals does not employ as wide a variety of scales as musical signals, we can conjecture that there may be other types of signals where this approach brings significant benefits, such as environmental sounds that distinguish themselves through a variety of time-scales. It should be noted as well that a similar set of features was successfully used in (Ravelli et al., 2008) for music genre recognition, again obtaining results comparable to state-of-the-art.

Finally, in this study we have only tried to extend the MFCC in its traditional use, but we do not take full advantage of the explanatory power of sparse decompositions. In particular, it is still an open issue whether our proposed features could be useful for instrument identification in a polyphonic, i.e., multi-instrumental, setting. This is perhaps where we will see a significant advantage in using sparse approximation over redundant transforms in computing and comparing these kinds of features.

Acknowledgements

The authors are grateful to E. Ravelli for help in computing aspects and S. Essid for sharing his database. M. Morvidone was partly supported by postdoctoral grants of the Ville de Paris and of the Université de Cergy-Pontoise. B. L. Sturm is supported by the Chateaubriand Fellowship No. 634146B. L. Daudet acknowledges partial support from the French Agence Nationale de la Recherche under contract ANR-06-JCJC-0027–01 DESAM.

References

Aucouturier, J.-J., Defréville, B., Pachet, F., 2007. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. J. Acoust. Soc. Amer. 122 (2), 881–891.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., Reynolds, D.A., 2004. A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process. 2004 (4), 430–451.

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M., 2008. Content-based music information retrieval: Current directions and future challenges. Proc. IEEE 96 (4), 668–696.

Chen, S.S., Donoho, D.L., Saunders, M.A., 1998. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20 (1), 33–61.

Couvreur, L., Laniray, M., 2004. Automatic noise recognition in urban environments based on artificial neural networks and hidden markov models. Proc. Internat. Congress Expo. Noise Control Eng. Prague, Czech Republic, pp. 1–8.

Cowling, M., Sitte, R., 2003. Comparison of techniques for environmental sound recognition. Pattern Recognition. Lett. 24 (15), 2895–2907.

Daudet, L., 2006. Sparse and structured decompositions of signals with the molecular matching pursuit. IEEE Trans. Audio, Speech, Lang. Process. 14 (5), 1808–1816.

Defréville, B., Roy, P., Rosin, C., Pachet, F., 2006. Automatic recognition of urban sound sources. Proc. Audio Eng. Soc. Paris, France, pp. 1–9.

Essid, S., 2005. Classification automatique des signaux audio-fréquence: Reconnaissance des instruments de musique. Ph.D. thesis, Université Pierre et Marie Curie, Paris 6.

Essid, S., Richard, G., David, B., 2006. Musical instrument recognition by pairwise classification strategies. IEEE Trans. Audio, Speech, Lang. Process. 14 (4), 1401–1412.

Ganchev, T., Fakotakis, N., Kokkinakis, G., 2005. Comparative evaluation of various MFCC implementations on the speaker verification task. Proc. Internat. Conf. on Speech Computer. Vol. 1. Patras, Greece, pp. 191–194.

Goto, M., Hashigushi, H., Nishimura, T., Oka, R., 2002. RWC music database: Popular, classical, and jazz music databases. Proc. Internat. Conf. on Music Information Retrieval. Paris, France, pp. 287–288.

Gribonval, R., Bacry, E., 2003. Harmonic decompositions of audio signals with matching pursuit. IEEE Trans. Signal Process. 51 (1), 101–111.

Herrera-Boyer, P., Peeters, G., Dubnov, S., 2003. Automatic classification of musical instrument sounds. J. New Music Research 32 (1), 3–21.

Hsu, C.W., Chang, C.C., Lin, C.J., 2009. A practical guide to support vector classification. Tech. Rep. http://www.csie.ntu.edu.tw/~cjlin, National Taiwan University, Taiwan, China.

IOWA, July 2009. University of Iowa musical instrument samples database. URL http://theremin.music.uiowa.edu/.

Joder, C., Essid, S., Richard, G., 2009. Temporal integration for audio classification with application to musical instrument classification. IEEE Trans. Acoust. Speech Signal Process. 17 (1), 174–184.

Krstulovic, S., Gribonval, R., 2006. MPTK: Matching pursuit made tractable. Proc. IEEE Internat. Conf. on Acoust., Speech Signal Process. Vol. 3. Toulouse, France, pp. 496–499.

Leveau, P., Vincent, E., Richard, G., Daudet, L., 2008. Instrument-specific harmonic atoms for mid-level music representation. IEEE Trans. Audio, Speech, Lang. Process 16 (1), 116–128.

Logan, B., 2000. Mel frequency cepstral coefficients for music modeling. Proc. Internat. Symp. Music Info. Retrieval. Plymouth, MA, pp. 1–13.

Mallat, S., Zhang, Z., 1993. Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. 41 (12), 3397–3415.

Manjunath, B.S., Salembier, P., Sikora, T. (Eds.), 2002. Introduction to MPEG-7: Multimedia Content Description Interface. J. Wiley, New York, NY.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice Hall, Upper Saddle River, New Jersey.

Ravelli, E., Richard, G., Daudet, L., 2008. Union of MDCT bases for audio coding. IEEE Trans. Audio, Speech, Lang. Proc 16 (8), 1361–1372.