

Aalborg Universitet



**AALBORG
UNIVERSITY**

Recursive nearest neighbor search in a sparse and multiscale domain for comparing audio signals

Sturm, Bob L.; Daudet, Laurent

Published in:
Signal Processing

DOI (link to publication from Publisher):
[10.1016/j.sigpro.2011.03.002](https://doi.org/10.1016/j.sigpro.2011.03.002)

Publication date:
2011

Document Version
Early version, also known as pre-print

[Link to publication from Aalborg University](#)

Citation for published version (APA):
Sturm, B. L., & Daudet, L. (2011). Recursive nearest neighbor search in a sparse and multiscale domain for comparing audio signals. *Signal Processing*, 91(12), 2836-2851. <https://doi.org/10.1016/j.sigpro.2011.03.002>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Recursive Nearest Neighbor Search in a Sparse and Multiscale Domain for Comparing Audio Signals

Bob L. Sturm (EURASIP# 7255)^{a,*}, Laurent Daudet (EURASIP # 2298)^b

^a*Department of Architecture and Media Technology, Aalborg University Copenhagen
Lautrupvang 15, 2750 Ballerup, Denmark*

^b*Institut Langevin (LOA), Université Paris Diderot – Paris 7,
UMR 7587, 10, rue Vauquelin, 75231 Paris, France*

Abstract

We investigate recursive nearest neighbor search in a sparse domain at the scale of audio signals. Essentially, to approximate the cosine distance between the signals we make pairwise comparisons between the elements of localized sparse models built from large and redundant multiscale dictionaries of time-frequency atoms. Theoretically, error bounds on these approximations provide efficient means for quickly reducing the search space to the nearest neighborhood of a given data; but we demonstrate here that the best bound defined thus far involving a probabilistic assumption does not provide a practical approach for comparing audio signals with respect to this distance measure. Our experiments show, however, that regardless of these non-discriminative bounds, we only need to make a few atom pair comparisons to reveal, e.g., the origin of an excerpted signal, or melodies with similar time-frequency structures.

Keywords: Multiscale decomposition; sparse approximation; time-frequency dictionary; audio similarity

1. Introduction

2 Sparse approximation is essentially the modeling of data with few terms
3 from a large and typically overcomplete set of atoms, called a “dictionary” [24].

*Corresponding author

Email addresses: boblsturm@gmail.com (Bob L. Sturm (EURASIP# 7255)),
laurent.daudet@espci.fr (Laurent Daudet (EURASIP # 2298))

4 Consider an $\mathbf{x} \in \mathbb{R}^K$, and a dictionary \mathcal{D} composed of N unit-norm atoms in the
5 same space, expressed in matrix form as $\mathbf{D} \in \mathbb{R}^{K \times N}$, where $N \gg K$. A pursuit
6 is an algorithm that decomposes \mathbf{x} in terms of \mathbf{D} such that $\|\mathbf{x} - \mathbf{D}\mathbf{s}\|^2 \leq \epsilon$ for
7 some error $\epsilon \geq 0$. (In this paper, we work in a Hilbert space unless otherwise
8 noted.) When \mathbf{D} is overcomplete, \mathbf{D} has full row rank and there exists an infinite
9 number of solutions to choose from, even for $\epsilon = 0$. Sparse approximation aims
10 to find a solution \mathbf{s} that is mostly zeros for ϵ small. In that case, we say that \mathbf{x}
11 is sparse in \mathcal{D} .

12 Matching Pursuit (MP) is an iterative descent sparse approximation method
13 based on greedy atom selection [17, 24]. We express the n th-order model of the
14 signal $\mathbf{x} = \mathbf{H}(n)\mathbf{a}(n) + \mathbf{r}(n)$, where $\mathbf{a}(n)$ is a length- n vector of weights, $\mathbf{H}(n)$
15 are the n corresponding columns of \mathbf{D} , and $\mathbf{r}(n)$ is the residual. MP augments
16 the n th-order representation, $\mathcal{X}_n = \{\mathbf{H}(n), \mathbf{a}(n), \mathbf{r}(n)\}$, according to:

$$\mathcal{X}_{n+1} = \left\{ \begin{array}{l} \mathbf{H}(n+1) = [\mathbf{H}(n) | \mathbf{h}_n], \\ \mathbf{a}(n+1) = [\mathbf{a}^T(n), \langle \mathbf{r}(n), \mathbf{h}_n \rangle]^T, \\ \mathbf{r}(n+1) = \mathbf{x} - \mathbf{H}(n+1)\mathbf{a}(n+1) \end{array} \right\} \quad (1)$$

17 using the atom selection criterion

$$\mathbf{h}_n = \arg \min_{\mathbf{d} \in \mathcal{D}} \|\mathbf{r}(n) - \langle \mathbf{r}(n), \mathbf{d} \rangle \mathbf{d}\|^2 = \arg \max_{\mathbf{d} \in \mathcal{D}} |\langle \mathbf{r}(n), \mathbf{d} \rangle| \quad (2)$$

18 where $\|\mathbf{d}\| = 1$ is implicit. The inner product here is defined $\langle \mathbf{x}, \mathbf{y} \rangle \triangleq \mathbf{y}^T \mathbf{x}$. This
19 criterion guarantees $\|\mathbf{r}(n+1)\|^2 \leq \|\mathbf{r}(n)\|^2$ [24]. Other sparse approximation
20 methods include orthogonal MP [28], orthogonal least squares (OLS) [41, 33],
21 molecular methods [9, 38, 19], cyclic MP and OLS [36], and minimizing a relaxed
22 sparsity measure [6]. These approaches have higher computational complexities
23 than MP, but can produce data models that are more sparse.

24 Sparse approximation is data-adaptive and can produce parametric and mul-
25 tiscale models having features that function more like mid-level “objects” than
26 low-level projections onto sets of vectors [9, 19, 38, 8, 22, 27, 32, 34, 43]. These
27 aspects make sparse approximation a compelling complement to state-of-the-art
28 approaches for and applications of comparing audio signals based upon, e.g.,

29 monoresolution cepstral and redundant time-frequency representations, such as
30 fingerprinting [42], cover song identification [26, 10, 3, 35], content segmentation,
31 indexing, search and retrieval [16, 4], artist or genre classification [39].

32 In the literature we find some existing approaches to working with audio sig-
33 nals in a sparse domain. Features built from sparse approximations can provide
34 competitive descriptors for music information retrieval tasks, such as beat track-
35 ing, chord recognition, and genre classification [40, 32]. Sparse representation
36 classifiers have been applied to music genre recognition [27, 5], and robust speech
37 recognition [12]. Parameters of sparse models can be compared using histograms
38 to find similar sounds in acoustic environment recordings [7, 8], or atoms can
39 be learned to compare and group percussion sounds [34]. Biologically-inspired
40 sparse codings of correlograms of sounds can be used to learn associations be-
41 tween descriptive high-level keywords and audio features such that new sounds
42 can be automatically categorized, and large collections of sounds can be queried
43 in meaningful ways [22]. Outside the realm of audio signals, sparse approxi-
44 mation has been applied to face recognition [44], object recognition [29], and
45 landmine detection [25].

46 In this paper, we discuss the comparison of audio signals in sparse domains,
47 but not specifically for fingerprinting or efficient audio indexing and search —
48 two tasks that have been convincingly solved [42, 13, 16, 18]. We explore the pos-
49 sibilities and effectiveness of comparing, atom-by-atom, audio signals modeled
50 using sparse approximation and large overcomplete time-frequency dictionaries.
51 Our contributions are three-fold: 1) we generalize an iterative nearest-neighbor
52 search algorithm to comparing subsequences [14, 15]; 2) we show that though
53 sparse models of audio signals can be compared by considering pairs of atoms,
54 the best bound so far derived [14, 15] does not make a practical procedure;
55 and 3) we show experimentally that the hierarchic comparison of audio signals
56 in a sparse domain still provides intriguing and informative results. Overall,
57 our work here shows that a sparse domain can facilitate comparisons of audio
58 signals in “hierarchical” ways through comparing individual elements of each
59 sparse data model organized roughly in order of importance.

60 In the next two sections, we discuss and elaborate upon a recursive method
 61 of nearest neighbor search in a sparse domain [14, 15]. We extend this method to
 62 comparing subsequences, and examine the practicality of probabilistic bounds
 63 on the distances between neighbors. In the fourth section, we describe several
 64 experiments in which we compare a variety of audio signals through comparisons
 65 of their sparse models. We conclude with a discussion about the results and
 66 several future directions.

67 2. Nearest Neighbor Search by Recursion in a Sparse Domain

68 Consider a set of signals

$$\mathcal{Y} \triangleq \{\mathbf{y}_i \in \mathbb{R}^K : \|\mathbf{y}_i\| = 1\}_{i \in \mathcal{I}} \quad (3)$$

69 where $\mathcal{I} = \{1, 2, \dots\}$ indexes this set, and a query signal $\mathbf{x}_q \in \mathbb{R}^K$, $\|\mathbf{x}_q\| = 1$.
 70 Assume that we have generated sparse approximations for all of these sig-
 71 nals $\hat{\mathcal{Y}} \triangleq \{\{\mathbf{H}_i(n_i), \mathbf{a}_i(n_i), \mathbf{r}_i(n_i)\} : \mathbf{y}_i = \mathbf{H}_i(n_i)\mathbf{a}_i(n_i) + \mathbf{r}_i(n_i)\}_{i \in \mathcal{I}}$ using a dic-
 72 tionary \mathbf{D} that spans the space \mathbb{R}^K , and giving the n_q -order representation
 73 $\{\mathbf{H}_q(n_q), \mathbf{a}_q(n_q), \mathbf{r}_q(n_q)\}$ for \mathbf{x}_q . Since \mathcal{D} spans \mathbb{R}^K , \mathcal{D} is “complete,” and any
 74 signal in \mathbb{R}^K is “compressible” in \mathcal{D} , meaning that we can order the represen-
 75 tation weights in $\mathbf{a}_i(n_i)$ or $\mathbf{a}_q(n_q)$ in terms of decreasing magnitude, i.e.,

$$0 < |[\mathbf{a}_i(n_i)]_{m+1}| \leq |[\mathbf{a}_i(n_i)]_m| \leq Cm^{-\gamma}, m = 1, 2, \dots, n_i - 1 \quad (4)$$

76 for n_i arbitrarily large, with $C > 0$, and where $[\mathbf{a}]_m$ is the m th element of the
 77 column vector \mathbf{a} . This can be seen in the magnitude representation weights in
 78 Fig. 1, which are weights of sparse representations of piano notes, described
 79 in Section 4.1. With MP and a complete dictionary, we are guaranteed $\gamma > 0$
 80 because $\|\mathbf{r}(n+1)\|^2 < \|\mathbf{r}(n)\|^2$ for all n [24].

81 Consider the Euclidean distance between two signals of the same dimension,
 82 which is the cosine distance for unit-norm signals. Thus, with respect to this
 83 distance, the $\mathbf{y}_i \in \mathcal{Y}$ nearest to \mathbf{x}_q is given by solving

$$\min_{i \in \mathcal{I}} \|\mathbf{y}_i - \mathbf{x}_q\| = \max_{i \in \mathcal{I}} \langle \mathbf{x}_q, \mathbf{y}_i \rangle. \quad (5)$$

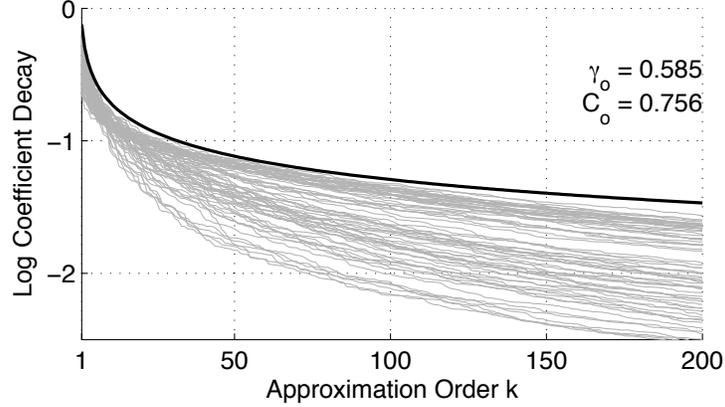


Figure 1: Gray lines show decays of representation weight magnitudes as a function of approximation order k for several decompositions of unit-norm signals (4-second recordings of single piano notes described in Section 4.1). Thick black line shows a global compressibility bound with its parameters.

84 We can express this inner product in terms of sparse representations

$$\begin{aligned} \langle \mathbf{x}_q, \mathbf{y}_i \rangle &= \langle \mathbf{H}_q(n_q) \mathbf{a}_q(n_q) + \mathbf{r}_q(n_q), \mathbf{H}_i(n_i) \mathbf{a}_i(n_i) + \mathbf{r}_i(n_i) \rangle \\ &= \mathbf{a}_i^T(n_i) \mathbf{H}_i^T(n_i) \mathbf{H}_q(n_q) \mathbf{a}_q(n_q) + O[\mathbf{r}_q, \mathbf{r}_i]. \end{aligned} \quad (6)$$

85 With a complete dictionary we can make $O[\mathbf{r}_q, \mathbf{r}_i]$ negligible by choosing ϵ ar-
86 bitrarily small, so we can express (5) as

$$\max_{i \in \mathcal{I}} \langle \mathbf{x}_q, \mathbf{y}_i \rangle \sim \max_{i \in \mathcal{I}} \mathbf{a}_i^T(n_i) \mathbf{G}_{iq} \mathbf{a}_q(n_q) = \max_{i \in \mathcal{I}} \sum_{m=1}^{n_i} \sum_{l=1}^{n_q} [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{ml} \quad (7)$$

87 where $[\mathbf{B} \bullet \mathbf{C}]_{ml} = [\mathbf{B}]_{ml} [\mathbf{C}]_{ml}$ is the Hadamard, or entry wise, product, $[\mathbf{B}]_{ml}$
88 is the element of \mathbf{B} in the m th row of the l th column, $\mathbf{G}_{iq} \triangleq \mathbf{H}_i^T(n_i) \mathbf{H}_q(n_q)$ is a
89 $n_i \times n_q$ matrix with elements from the Gramian of the dictionary, i.e., $\mathbf{G} \triangleq \mathbf{D}^T \mathbf{D}$,
90 and finally we define the outer product of the weights

$$\mathbf{A}_{iq} \triangleq \mathbf{a}_i(n_i) \mathbf{a}_q^T(n_q). \quad (8)$$

91 2.1. Recursive Search Limited by Bounds

92 Since we expect the decay of the magnitude of elements in $\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}$ to be
93 fastest in diagonal directions by (4), we define a recursive sum along the M

94 anti-diagonals starting at the top left:

$$S_{iq}(M) \triangleq S_{iq}(M-1) + \sum_{m=1}^M [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{m(M-m+1)} \quad (9)$$

95 for $M = 2, 3, \dots, \min(n_i, n_q)$, and setting $S_{iq}(1) = [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{11}$. With this we
 96 can express the argument of (7) as

$$\langle \mathbf{x}_q, \mathbf{y}_i \rangle \approx \sum_{m=1}^{n_i} \sum_{l=1}^{n_q} [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{ml} = S_{iq}(M) + R(M) \quad (10)$$

97 where at step M , we are comparing M additional pairs of atoms to those con-
 98 sidered in the previous steps. $R(M)$ is a remainder that we will bound. The
 99 total number of atom pairs contributing to $S_{iq}(M)$ (9) is

$$P(M) \triangleq \sum_{m=1}^M m = M(M+1)/2. \quad (11)$$

100 The approach taken by Jost et al. [14, 15] to find the nearest neighbors of
 101 \mathbf{x}_q in \mathcal{Y} bounds the remainder $R(M)$ by compressibility (4). Assuming we have
 102 a positive upper bound on the remainder, i.e., $R(M) \leq \tilde{R}(M)$. we know lower
 103 and upper bounds on the cosine distance $L_{iq}(M) \leq \langle \mathbf{x}_q, \mathbf{y}_i \rangle \leq U_{iq}(M)$, where

$$L_{iq}(M) \triangleq S_{iq}(M) - \tilde{R}(M) \quad (12)$$

$$U_{iq}(M) \triangleq S_{iq}(M) + \tilde{R}(M). \quad (13)$$

104 Finding elements of \mathcal{Y} close to \mathbf{x}_q with respect to (5) can be done recursively
 105 over the approximation order M . For a given M , we find $\{S_{iq}(M)\}_{i \in \mathcal{I}}$, com-
 106 pute the remainder $\tilde{R}(M)$, and eliminate signals that are not sufficiently close
 107 to \mathbf{x}_q with respect to their cosine distance by comparing the bounds. This
 108 approach is similar to hierarchical ones, e.g., [21], where the features become
 109 more discriminable as the search process runs. (Also note that compressibility
 110 is similar to the argument made in justifying the truncation of Fourier series in
 111 early work on similarity search [1, 11, 30], i.e., that power spectral densities of
 112 many time-series decay like $\mathcal{O}(|f|^{-b})$ with $b > 1$.)

113 Starting with $M = 1$, we compute the sets $\{L_{iq}(1)\}_{i \in \mathcal{I}}$ and $\{U_{iq}(1)\}_{i \in \mathcal{I}}$,
 114 that is, the first-order upper and lower bounds of the set of distances of \mathbf{x}_q

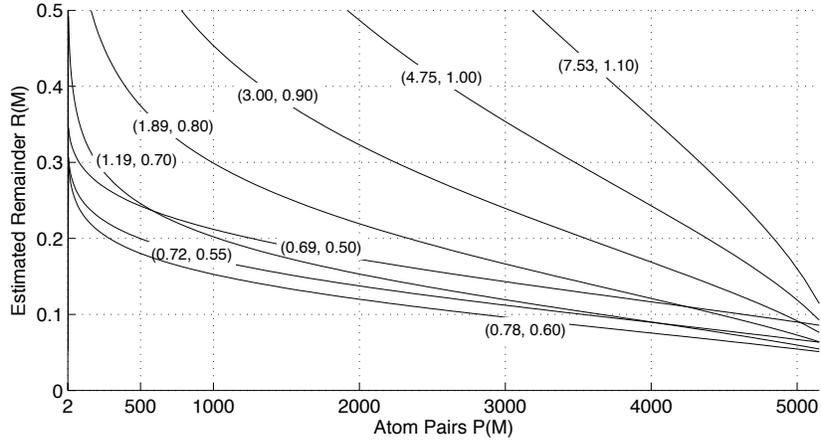


Figure 2: Estimated remainder (assuming unit-norm signals) using bound in (16) with $p = 0.2$ (probability that remainder does not exceed bound) and $n = 100$ (number of elements in each sparse model) as a function of the number of atom pairs already considered for several pairs of compressibility parameters (C, γ) estimated from the dataset used to produce Fig. 1.

115 from all signals in \mathcal{Y} . Then we find the index of the largest lower bound
 116 $i_{\max} = \arg \max_{i \in \mathcal{I}} L_{iq}(1)$, and reduce the search space to $\mathcal{I}_1 \triangleq \{i \in \mathcal{I} : U_{iq}(1) \geq$
 117 $L_{i_{\max}q}(1)\}$, since all other data have a least upper bound on their inner prod-
 118 uct with \mathbf{x}_q than the greatest lower bound in the set. For the next step, we
 119 compute the sets $\{L_{iq}(2)\}_{i \in \mathcal{I}_1}$ and $\{U_{iq}(2)\}_{i \in \mathcal{I}_1}$, find the index of the maxi-
 120 mum $i_{\max} = \arg \max_{i \in \mathcal{I}_1} L_{iq}(2)$, and construct the reduced set $\mathcal{I}_2 \triangleq \{i \in \mathcal{I}_1 :$
 121 $U_{iq}(2) \geq L_{i_{\max}q}(2)\}$. Continuing in this way, we find the elements of \mathcal{Y} closest
 122 to \mathbf{x}_q at each M with respect to the cosine distance by recursing into the sparse
 123 approximations of the signals.

124 2.2. Bounding the Remainder

125 To reduce the search space quickly we desire that (12) and (13) converge
 126 quickly to the neighborhood of $\langle \mathbf{x}_q, \mathbf{y}_i \rangle$, or in other words, that the bounds on
 127 the remainder quickly become discriminative. Jost et al. [14, 15] derive three
 128 different bounds on $R(M)$. From the weakest to the strongest, these are:

129 1. $[\mathbf{G}_{iq}]_{ml} = 1$ (worst case scenario, and impossible for $n > 1$)

$$R(M) \leq C^2 (\|\mathbf{c}_M^\gamma\|_1 + \|\mathbf{d}^\gamma\|_1) \quad (14)$$

130 2. $[\mathbf{G}_{iq}]_{ml} \sim \text{iid Bernoulli}(0.5)$, $\Omega = \{-1, 1\}$ (impossible for $n > 1$)

$$R(M) \leq C^2 \sqrt{\ln 4} (\|\mathbf{c}_M^\gamma\|_2^2 + \|\mathbf{d}^\gamma\|_2^2)^{1/2} \quad (15)$$

131 3. $[\mathbf{G}_{iq}]_{ml} \sim \text{iid Uniform}$, $\Omega = [-1, 1]$,

$$R(M) \leq C^2 \sqrt{2/3} \text{Erf}^{-1}(p) (\|\mathbf{c}_M^\gamma\|_2^2 + \|\mathbf{d}^\gamma\|_2^2)^{1/2} \quad (16)$$

132 with probability $0 \leq p \leq 1$

133 where we define the following vectors for $n \triangleq \min(n_i, n_q)$ and $M = 2, \dots, n$

$$\mathbf{c}_M^\gamma \triangleq \{[l(m-l+1)]^{-\gamma} : m = M+1, \dots, n; l = 1, \dots, m\} \quad (17)$$

$$\mathbf{d}^\gamma \triangleq \{[l(n-m+1)]^{-\gamma} : m = 1, \dots, n-1; l = m+1, \dots, n\}. \quad (18)$$

134 Appendix A gives derivations of these bounds, as well as the efficient computa-
 135 tion of (16) for the special case of $\gamma = 0.5$. The parameters (C, γ) describe the
 136 compressibility of the signals in the dictionary (4). The bounds of (15) and (16)
 137 are much more discriminative than (14) because they involve an ℓ_2 -norm at the
 138 price of uncertainty in the bound. The bound in (16) is attractive because we
 139 can tune it with the parameter p , which is the probability that the remainder
 140 will not exceed the bound. Figure 2 shows bounds based on (16) for several
 141 pairs of compressibility parameters for the dataset used to produce Fig. 1.

142 2.3. Estimating the Compressibility Parameters

143 The bounds (14)–(16), and consequently the number of atom pairs we must
 144 consider before the bounds become discriminable, depend on the compressibility
 145 parameters, (C, γ) — which themselves depend in a complex way on the signal,
 146 the dictionary, and the method of sparse approximation. Figure 3 shows the
 147 error surface, feasible set, and the optimal parameters for the dataset used to
 148 produce Fig. 1. We describe our parameter estimation procedure in Appendix
 149 B. The resulting bound is shown in black in Fig. 1. These compressibility
 150 parameters also agree with those seen in Fig. 2.

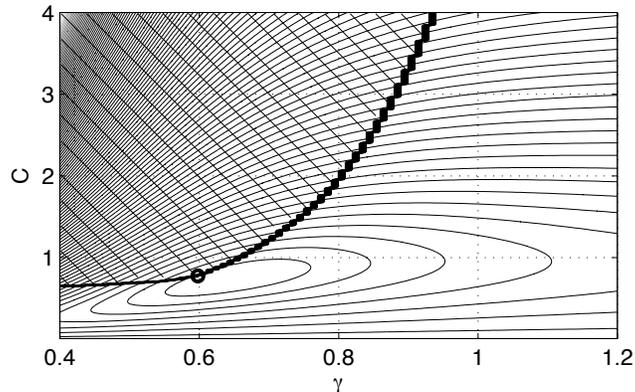


Figure 3: Error surface as a function of the compressibility parameters for the dataset used to produce Fig. 1, with the feasible set shaded at top left, and optimal parameters marked by a circle.

151 **3. Subsequence Nearest Neighbor Search in a Localized Sparse Do-**
 152 **main**

153 The recursive nearest neighbor search so far described has the obvious limi-
 154 tation that it cannot be applied to comparing subsequences of large data vectors,
 155 as is natural for comparing audio signals. Thus, we must adapt its structure to
 156 work for comparing subsequences in a set of data

$$\mathcal{Y} \triangleq \{\mathbf{y}_i \in \mathbb{R}^{N_i} : N_i \geq K\}_{i \in \mathcal{I}} \quad (19)$$

157 (note that now we do not restrict the norms of these signals). We can create
 158 from the elements of \mathcal{Y} a new set of all subsequences having the same length as
 159 a K -dimensional query \mathbf{x}_q ($K < N_i$):

$$\mathcal{Y}_K \triangleq \left\{ \mathbf{P}_t \mathbf{y}_i / \|\mathbf{P}_t \mathbf{y}_i\| : t \in \mathcal{T}_i = \{1, 2, \dots, N_i - K + 1\}, \mathbf{y}_i \in \mathcal{Y} \right\} \quad (20)$$

160 where \mathbf{P}_t extracts a K -length subsequence in \mathbf{y}_i starting a time-index t (it is an
 161 identity matrix of size K starting a column t in a $K \times N_i$ matrix of zeros). The set
 162 \mathcal{T}_i are times at which we create length- K subsequences from \mathbf{y}_i . If we decompose
 163 each of these by sparse approximation, then we can use the framework in the
 164 previous section. However, sparse approximation is an expensive operation that

165 we want to do only once for the entire signal, and independent of the length of
 166 \mathbf{x}_q .

167 To address this problem, we instead approximate each element in \mathcal{Y}_K by
 168 building local sparse representations from the global sparse approximations of
 169 each \mathbf{y}_i , and then calculating their distance to \mathbf{x}_q using the framework in the
 170 previous section. From here on we consider only the K -length subsequences of
 171 a single element $\mathbf{y}_i \in \mathcal{Y}$ without loss of generality (i.e., all other elements of
 172 \mathcal{Y} can be included as subsequences). Toward this end, consider that we have
 173 decomposed the N_i -length signal \mathbf{y}_i using a complete dictionary to produce the
 174 representation $\{\mathbf{H}_i(n_i), \mathbf{a}_i(n_i), \mathbf{r}_i(n_i)\}$. From this we construct the local sparse
 175 representations of \mathbf{y}_i :

$$\widehat{\mathcal{Y}}_K \triangleq \left\{ \{\mathbf{P}_t \mathbf{H}_i(n_i), \xi_t \mathbf{a}_i(n_i), \mathbf{P}_t \mathbf{r}_i(n_i)\} : t \in \mathcal{T}_i \right\} \quad (21)$$

176 where the time partition \mathcal{T}_i is the set of all times at which we extract a K -
 177 length subsequence from \mathbf{y}_i , and ξ_t is set such that $\|\xi_t \mathbf{P}_t \mathbf{y}_i\| = 1$, i.e., each
 178 length- K subsequence is unit-norm. For each K -dimensional subsequence, (7)
 179 now becomes

$$\begin{aligned} \max_{t \in \mathcal{T}_i} \langle \mathbf{x}_q, \mathbf{P}_t \mathbf{y}_i \rangle &= \max_{t \in \mathcal{T}_i} \left[\langle \mathbf{H}_q(n_q) \mathbf{a}_q(n_q), \xi_t \mathbf{P}_t \mathbf{H}_i(n_i) \mathbf{a}_i(n_i) \rangle + O[\mathbf{r}_q, \mathbf{r}_i] \right] \\ &\approx \max_{t \in \mathcal{T}_i} \xi_t \mathbf{a}_i^T(n_i) \mathbf{H}_i^T(n_i) \mathbf{P}_t^T \mathbf{H}_q(n_q) \mathbf{a}_q(n_q) \\ &= \max_{t \in \mathcal{T}_i} \xi_t \sum_{m=1}^{n_i} \sum_{l=1}^{n_q} [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}(t)]_{ml} \end{aligned} \quad (22)$$

180 where \mathbf{A}_{iq} is defined in (8), we define the time-localized Gramian

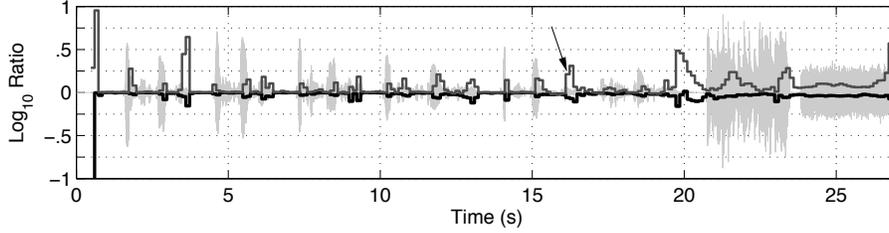
$$\mathbf{G}_{iq}(t) \triangleq \mathbf{H}_i^T(n_i) \mathbf{P}_t^T \mathbf{H}_q(n_q) \quad (23)$$

181 and we have excluded the terms involving the residuals because we can make
 182 them arbitrarily small.

183 3.1. Estimating the Localized Energy

184 The only thing left to do is find an expression for ξ_t so that each subse-
 185 quence is comparable with the others with respect to the cosine distance. We

(a) Six Speech Signals (0-20 s), Music Excerpt (21-23 s), Realization of GWN (24-27 s)



(b) Music: Orchestra

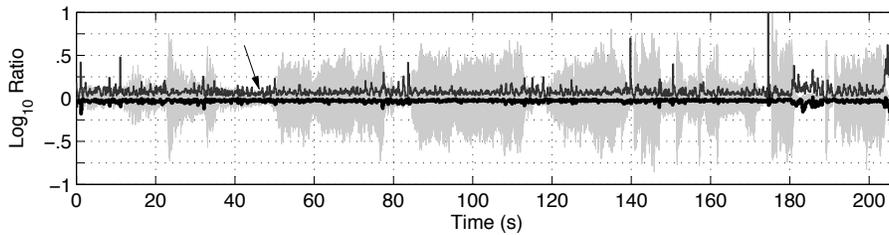


Figure 4: Short-term energy ratios, $\log_{10}(\sum_{j=1}^{n_t} w_j a_j^2 / \|\mathbf{P}_t \mathbf{y}_i\|^2)$, over 1 s windows (hopped 125 ms) for MP decompositions using 8xMDCT [31] to a global residual energy 30 dB below the initial signal energy. Arrow points to line (top, gray) using weighting $w_j = 1$. The other line (bottom, black) uses (25). Data in (a) are described in Section 4.2; data in (b) are described in Section 4.3.

186 assume that the localized energy can be approximated from the local sparse
 187 representation in the following way assuming $\|\mathbf{P}_t \mathbf{y}_i\| > 0$

$$\xi_t = \|\mathbf{P}_t \mathbf{y}_i\| \approx \sqrt{\mathbf{a}_i^T(n_i) \mathbf{H}_i^T(n_i) \mathbf{P}_t^T \mathbf{P}_t \mathbf{H}_i(n_i) \mathbf{a}_i(n_i)} \approx \sqrt{\sum_{j=1}^{n_t} w_j a_j^2} \quad (24)$$

188 where the n_t weights $a_j \in \{[\mathbf{a}_i(n_i)]_m : [\mathbf{H}_i^T(n_i) \mathbf{P}_t^T \mathbf{P}_t \mathbf{H}_i(n_i)]_{ml} \neq 0, 1 \leq m, l \leq$
 189 $n_i\}$ are those associated with atoms having support in $[t, t + K)$, and w_j we
 190 define to weigh the contribution of a_j^2 to the localized energy estimate. We set
 191 $\xi_t = 0$ if $\sum_{j=1}^{n_t} a_j^2 = 0$.

192 If all atoms contributing to the subsequence have their entire support in
 193 $[t, t + K)$, and are orthonormal, then we can set each $w_j = 1$. This does not
 194 hold for subsequences of a signal decomposed using an overcomplete dictionary,
 195 as shown by Fig. 4. For much of the time we see $\sum_{j=1}^{n_t} a_j^2 \geq \|\mathbf{P}_t \mathbf{y}_i\|^2$, which

196 means our localized estimate of the segment energy is greater than its real value.

197 This will make ξ_t and consequently (22) smaller.

198 Instead, we make a more reasonable estimate of $\|\mathbf{P}_t \mathbf{y}_i\|$ by accounting for
 199 the fact that atoms can have support outside $[t, t + K)$. For instance, if an atom
 200 has some fraction of support in the subsequence we multiply its weight by that
 201 fraction. We thus weigh the contribution of the j th atom to the subsequence
 202 norm using

$$w_j = \begin{cases} 1, & u_j \geq t, u_j + s_j \leq t + K \\ (K/s_j)^2, & u_j < t, u_j + s_j \geq t + K \\ (u_j + s_j - t)^2/s_j^2, & u_j < t, t < u_j + s_j \leq t + K \\ (t + K - u_j)^2/s_j^2, & t \leq u_j < t + K, u_j + s_j > t + K \end{cases} \quad (25)$$

203 where u_j and s_j are the position and scale, respectively, of the atom associated
 204 with the weight a_j . In other words, if an atom is completely in $[t, t + K)$, it
 205 contributes all of its energy to the approximation; otherwise, it contributes only
 206 a fraction based on how its support intersects $[t, t + K)$. With this we are now
 207 slightly underestimating the localized energies, as seen in Fig. 4. In both of
 208 these cases for $\{w_j\}$, however, we can assume by the energy conservation of MP
 209 [24] that as the subsequence length becomes larger our error in estimating the
 210 subsequence energy goes to zero, i.e.,

$$\lim_{K \rightarrow N_i} \|\mathbf{P}_t \mathbf{y}_i\|^2 - \sum_{j=1}^{n_i} w_j a_j^2 = \|\mathbf{P}_t \mathbf{r}_i(n_i)\|^2. \quad (26)$$

211 With a complete dictionary, the right hand side can be made zero. Signifi-
 212 cant departures from the energy estimate of subsequences can be due to the
 213 interactions between atoms [37].

214 3.2. Recursive Subsequence Search Limited by Bounds

215 Now, similar to (9) and (10), we can say,

$$\langle \mathbf{x}_q, \mathbf{P}_t \mathbf{y}_i \rangle \approx \xi_t \sum_{m=1}^{n_t} \sum_{l=1}^{n_q} [\mathbf{A}_{iq}(t) \bullet \mathbf{G}_{iq}(t)]_{ml} = S_{iq}(t, M) + R(t, M) \quad (27)$$

216 where for $M = 2, 3, \dots, \min(n_i, n_q)$, and with $S_{iq}(t, 1) = \xi_t[\mathbf{A}_{iq}(t) \bullet \mathbf{G}_{iq}(t)]_{11}$,

$$S_{iq}(t, M) \triangleq S_{iq}(t, M - 1) + \xi_t \sum_{m=1}^M [\mathbf{A}_{iq}(t) \bullet \mathbf{G}_{iq}(t)]_{m(M-m+1)}. \quad (28)$$

217 The problem of finding the subsequence closest to \mathbf{x}_q with respect to the co-
 218 sine distance can now be done iteratively over M by bounding each remainder
 219 $R(t, M)$ using (14), (15), or (16), and the method presented in in Section 2.1.
 220 Furthermore, we can compare only a subset of all possible subsequences using
 221 a coarser time partition \mathcal{T}_i .

222 3.3. Practicality of the Bounds for Audio Signals

223 The experiments by Jost et al. [14, 15] use small images (128 square) and
 224 orthogonal wavelet decompositions, which do not translate to audio signals de-
 225 composed over redundant time-frequency dictionaries. Jost et al. [14, 15] do
 226 not state the compressibility parameters they use, but for the high-dimensional
 227 audio signals with which we work in this paper it is not unusual to have $\gamma \approx 0.5$
 228 when using MP and highly overcomplete dictionaries. We find that decom-
 229 posing 4 s segments of music signals (single channel, 44.1 kHz sample rate,
 230 representation weights shown in Fig. 1) using the dictionary in Table 1 requires
 231 on average 2,375 atoms to reduce the residual energy 20 dB below the initial
 232 signal energy. Thus, for the bound (16) using $n = 2,375$ atoms, and with the
 233 parameters $(C, \gamma) = (0.4785, 0.5)$ (in the feasible set), Fig. 5 clearly shows that
 234 in order to have any discriminable bound (say ± 0.2 for unit-norm signals) we
 235 must either select a low value for p — in which case we are assuming the first
 236 atom comparison is approximately the cosine distance — or we must make over
 237 a million pairwise comparisons.

238 This is not practical for signals of large dimension, and dictionaries contain-
 239 ing billions of time-frequency atoms. There is no possibility of tabulating the
 240 dictionary Gramian for quick lookup of atom pair correlations; and the cost of
 241 looking up atoms in million-atom decompositions is expensive as well. It is clear
 242 then that the tightest bound given in (16) is not practical for efficiently discrim-

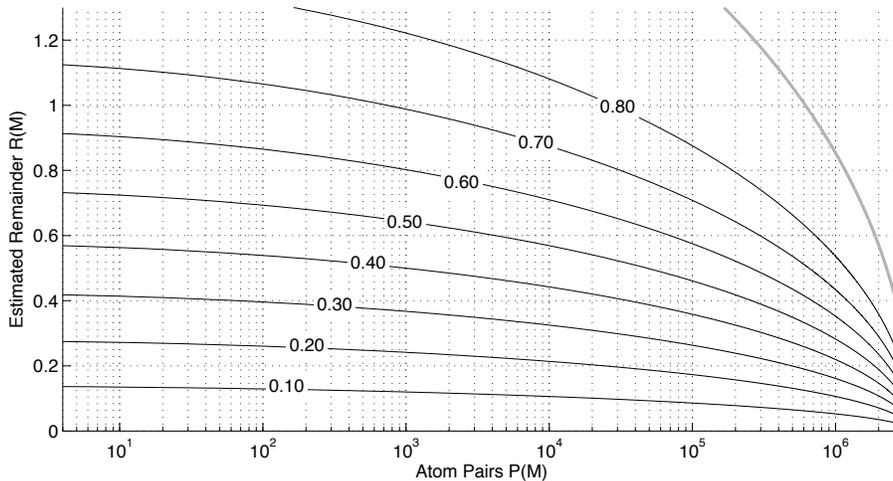


Figure 5: Estimated remainder (assuming unit-norm signals) as a function of the number of atom pairs already considered for dataset used to produce Fig. 1. Gray: bound in (15). Black, numbered: bound in (16) for several labeled p (probability that remainder does not exceed bound) with $n = 2,375$ (number of elements in each sparse model), and $(C, \gamma) = (0.4785, 0.5)$.

243 inating distances between audio signals with respect to their cosine distance (5)
 244 decomposed by MP and time-frequency dictionaries.

245 4. Experiments in Comparing Audio Signals in a Sparse Domain

246 Though approximate nearest neighbor subsequence search of sparsely ap-
 247 proximated audio signals with the bound (16) is impractical, we have found
 248 that approximating the cosine distance in a sparse domain has some intrigu-
 249 ing behaviors. We now present several experiments where we compare different
 250 types of audio data in a sparse domain under a variety of conditions. All signals
 251 are single channel, and have a sampling rate of 44.1 kHz. We decompose each
 252 by MP [17] using either the dictionary in Table 1, or the 8xMDCT dictionary
 253 [31].

254 4.1. Experiment 1: Comparing Piano Notes

255 In this experiment, we look at how well low-order sparse approximations
 256 of sampled piano notes embody their harmonic characteristics by comparing

s (samples/ms)	Δ_u (samples/ms)	Δ_f (Hz)
128/3	32/0.7	43.1
256/6	64/2	43.1
512/12	128/3	43.1
1024/23	256/6	43.1
2048/46	512/12	21.5
4096/93	1024/23	10.8
8192/186	2048/46	5.4
16384/372	4096/93	2.7
32768/743	8192/186	1.3

Table 1: Time-frequency dictionary parameters (44.1 kHz sampling rate): atom scale s , time resolution Δ_u , and frequency resolution Δ_f . Finer frequency resolution for small-scale atoms is achieved with interpolation by zero-padding.

257 them using the methods presented in Section 2. The data in set ‘A’ are 68 notes
258 (chromatically spanning A0 to G#6) on a real and somewhat in-tune piano; and
259 in set the data ‘B’ are 39 notes (roughly a C major scale C0 to D6) on a real
260 and very out-of-tune piano with very poor recording conditions. We truncate
261 all signals to have a dimension of 176,400 (4 seconds), and decompose each
262 by MP [17] over a redundant dictionary of time-frequency Gabor atoms, the
263 parameters of which are summarized in Table 1. We stop each decomposition
264 once its residual energy drops 40 dB below the initial energy. We normalize the
265 weights of each model by the square root energy of the respective signal. We do
266 not align the time-domain signals such that the note onsets occur at the same
267 time. Figure 1 shows the ordered decays of the weights in the sparse models of
268 data set ‘A’.

269 Figure 6(a) shows the magnitude correlations between all pairs of signals in
270 set ‘A’ evaluated in the time-domain. The overtone series is clear as diagonal
271 lines offset at 12, 19, 24, and 28, semitones from the main diagonal. Figure 6(b)
272 shows the approximated magnitude correlations (9) using only $M = 10$ atoms
273 from each signal approximation (thus $P(10) = 55$ atom pairs). Even though the

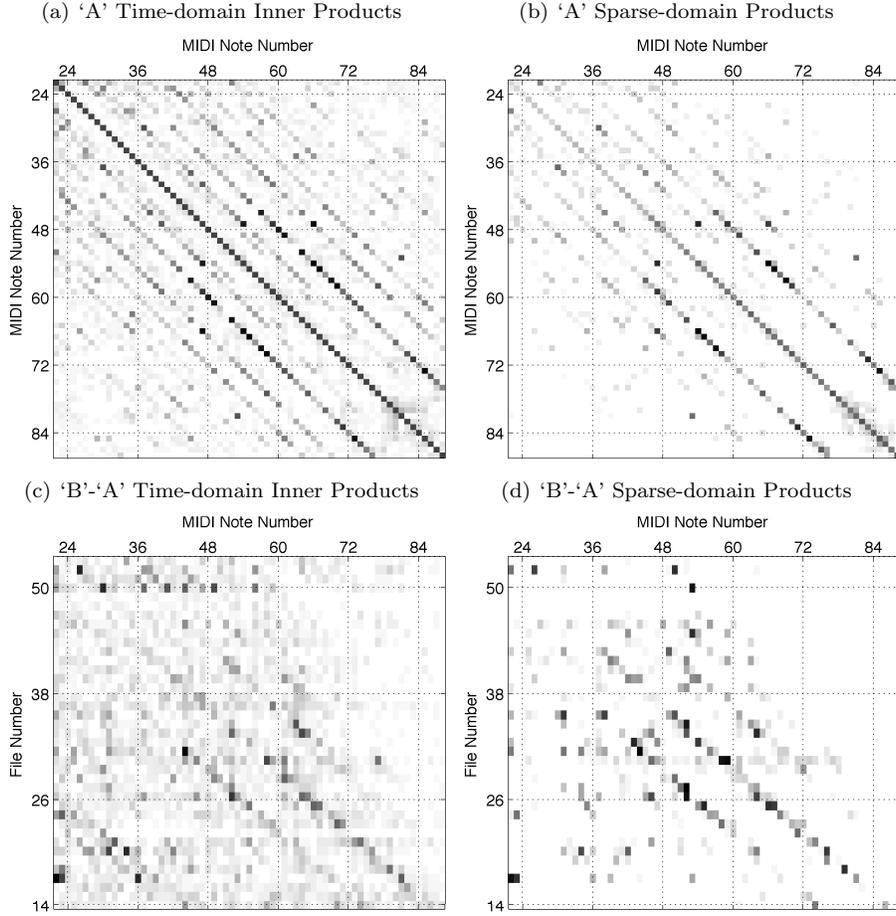


Figure 6: $|S_{iq}(10)|$ (9) for two sets of recorded piano notes. (a) and (b): Set ‘A’ compared with itself in time and sparse domains ($M = 10$). (c) and (d): Set ‘B’ (rows) compared with set ‘A’ (columns) in time and sparse domains ($M = 10$). Elements on main diagonals in (a) and (b) are scaled by 0.25 to increase contrast of other elements.

274 mean number of atoms in this set of models is about 7000 we still see portions
 275 of the overtone series. The diagonal in Fig. 6(b) does not have a uniform
 276 color because low notes have longer sustain times than high notes, and the
 277 sparse models thus have more time-frequency atoms with greater energies spread
 278 over the support of the signal. Figure 6(c) show the magnitude correlations
 279 between sets ‘B’ and ‘A’ evaluated in the time-domain; and Fig. 6(d) shows the

280 magnitude correlations (9) using only $M = 10$ atoms from each model. In a
 281 sparse domain, we can more clearly see the relationships between the two sets
 282 because the first 10 terms of each model are most likely related to the stable
 283 harmonics of the notes, and not to the noise. We can see a diatonic scale starting
 284 around MIDI number 36, as well as the fact that the pitch scale in data set ‘B’
 285 lies somewhere in-between the semitones in data set ‘A’.

286 Figure 7(a) shows the approximate magnitude correlations $|S_{iq}(M)|$ (9), as
 287 well as the upper and lower bounds on the remainder using the tightest bound
 288 (16) with $p = 0.2$, and $n = 100$, for the signal A3 from set ‘A’ and the rest of the
 289 set. Here we can see that the lower bound for the largest magnitude correlation
 290 exceeds the upper bound of all the rest after comparing only $M = 19$ atoms
 291 from each decomposition. All but five of the signals can be excluded from the
 292 search after $M = 6$. The four other signals having the largest approximate
 293 magnitude correlation are labeled, and are harmonically related to the signal
 294 through its overtone series. With a signal selected from set ‘B’ and compared to
 295 set ‘A’, Fig. 7(b) shows that we must compare many more atoms between the
 296 models until the bounds have any discriminability. After $P(M) = 1500$ atom
 297 comparisons we can see that the largest magnitudes $|S_{iq}(M)|$ (9) are roughly
 298 harmonically related to the detuned D5 from set ‘B’.

299 As a final part of this experiment, we look at the effects of comparing atoms
 300 with parameters within some subset. As done in Fig. 6(d), we compare the
 301 sparse approximations of two different sets of piano notes, but here we only
 302 consider those atoms that have scales greater than 186 ms. This in effect means
 303 that we look for signals that share the same “long-term” time-frequency be-
 304 haviors. The resulting $|S_{iq}(10)|$ (9) is shown in Fig. 8. We see correlations
 305 between the notes much more clearly compared with Fig. 6(d). Removing the
 306 short-term phenomena improves “tonal”-level comparisons between the signals
 307 because non-overlapping yet energetic short atoms are replaced by atoms rep-
 308 resentative of the note harmonics.

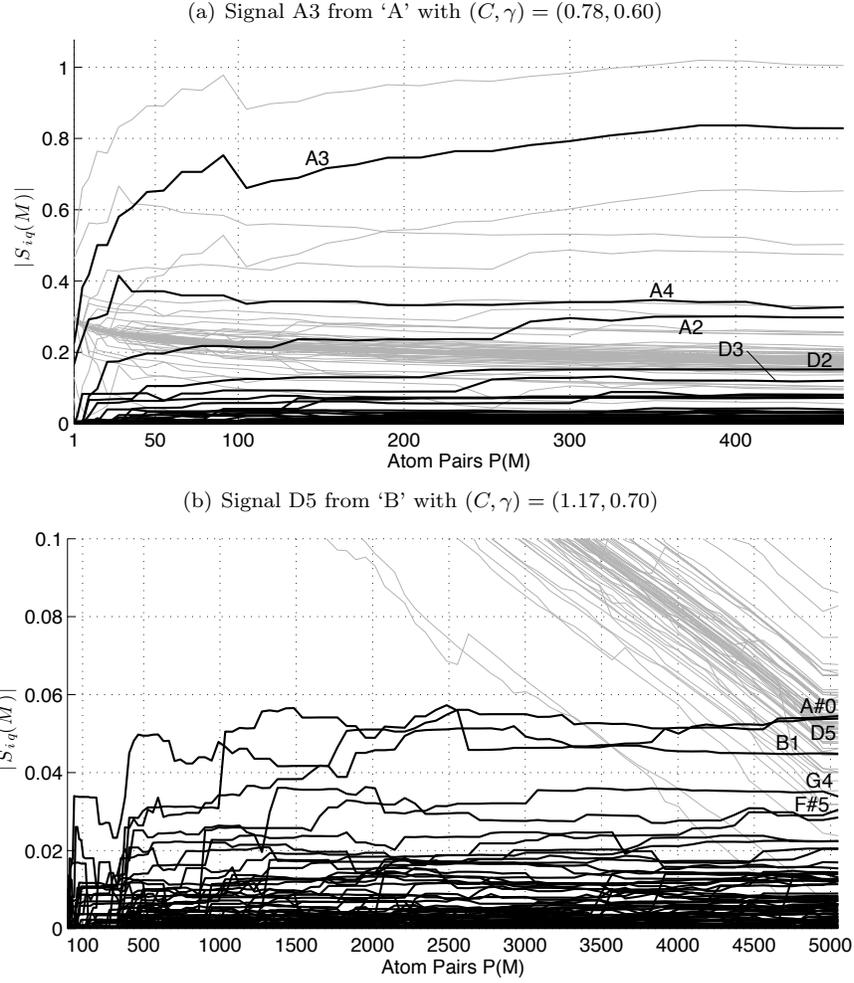


Figure 7: Black: $|S_{iq}(M)|$ (9) as a function of the number of atom pairs considered for the set of piano notes in 'A' with a signal from either (a) 'A' (note A3) or (b) 'B' (note D5 approximately). Gray: for each $S_{iq}(M)$, magnitudes of $L_{iq}(M)$ (12) and $U_{iq}(M)$ (13) using bound in (16) with $p = 0.2$ (probability that remainder does not exceed bound), and $n = 100$ (number of elements in each sparse model). Largest magnitude correlations are labeled. Note differences in axes.

309 *4.2. Experiment 2: Comparing Speech Signals*

310 In this experiment, we test how efficiently using (28) we can find in a speech
 311 signal the time from which we extract some \mathbf{x}_q . We also test how distortion in
 312 the query affects these results. We make a signal by combining six segments of

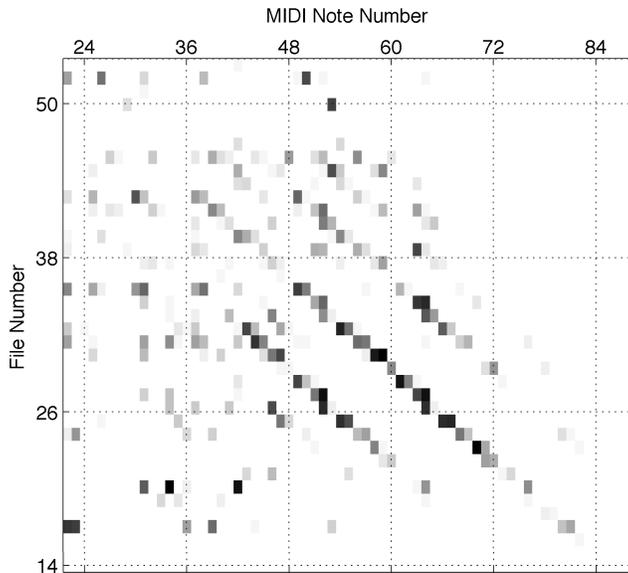


Figure 8: $|S_{iq}(10)|$ (9) for two sets of recorded piano notes in a sparse domain using only the atoms with duration at least 186 ms. Compare with Fig. 6(d).

313 speech, a short music segment, and white noise, shown in Fig. 4(a). The six
 314 speech segments are the same phrase spoken by three females and three males:
 315 “Cottage cheese with chives is delicious.” We extract from one of these speech
 316 signals the word “cheese,” to create \mathbf{x}_q with duration of 603 ms, shown at top in
 317 Fig. 9. We decompose this signal using MP and the 8xMDCT dictionary [31].

318 We distort the query in two ways: with additive WGN (AWGN), and with
 319 an interfering sound having a high correlation with the dictionary. In the first
 320 case, shown in the middle in Fig. 9, the signal $\mathbf{x}'_q = (\alpha\mathbf{x}_q + \mathbf{n})/|\alpha\mathbf{x}_q + \mathbf{n}|$ is
 321 the original \mathbf{x}_q distorted by a unit-norm AWGN signal \mathbf{n} . We set $\alpha = 0.3162$
 322 such that $10 \log_{10}(|\alpha\mathbf{x}_q|^2/|\mathbf{n}|^2) = 20 \log_{10}(|\alpha|) = -10$ dB. For this signal,
 323 we find the following statistics from 10,000 realizations of the AWGN signal:
 324 $E[|\langle \mathbf{x}_q, \mathbf{n} \rangle|] \approx 1 \times 10^{-5}$, $\text{Var}[|\langle \mathbf{x}_q, \mathbf{n} \rangle|] \approx 4 \times 10^{-6}$. We also find the fol-
 325 lowing statistics for the 8xMDCT dictionary: $E[\max_{\mathbf{d} \in \mathcal{D}} |\langle \mathbf{n}, \mathbf{d} \rangle|] \approx 5 \times 10^{-4}$,
 326 $\text{Var}[\max_{\mathbf{d} \in \mathcal{D}} |\langle \mathbf{n}, \mathbf{d} \rangle|] \approx 2 \times 10^{-5}$. Thus, the noise signal is not well-correlated
 327 either with the original signal or the 8xMDCT dictionary. In the second case,

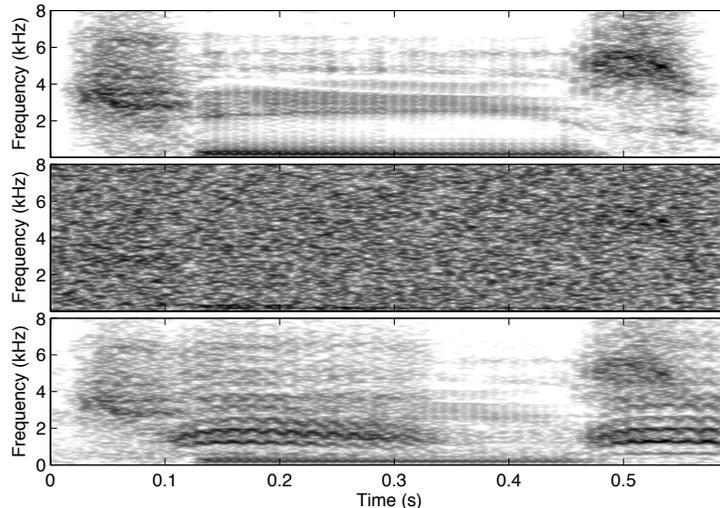


Figure 9: Log spectrograms of the query signals with which we search. Top: query of male saying “cheese.” Middle: query distorted with additive white Gaussian noise (AWGN) with SNR = -10 dB. Bottom: query distorted with interfering crow sound with SNR = -5 dB.

328 shown at the bottom of Fig. 9, we distort the signal by adding the sound of
 329 a crow \mathbf{c} so that $\mathbf{x}'_q = (\alpha\mathbf{x}_q + \mathbf{c})/||\alpha\mathbf{x}_q + \mathbf{c}||_2$ with $||\mathbf{c}|| = 1$. Here, we set
 330 $\alpha = 0.5623$ given by $20 \log_{10}(|\alpha|) = -5$ dB. For this interfering signal, we find
 331 that $|\langle \mathbf{x}_q, \mathbf{c} \rangle| \approx 2 \times 10^{-3}$, but $\max_{\mathbf{d} \in \mathcal{D}} |\langle \mathbf{c}, \mathbf{d} \rangle| \approx 0.21$, which is higher than
 332 $\max_{\mathbf{d} \in \mathcal{D}} |\langle \mathbf{x}_q, \mathbf{d} \rangle| \approx 0.17$. In this case, unlike for the AWGN interference, it is
 333 likely that the sparse approximation of the signal with the crow interference will
 334 have atoms in its low-order model due to the crow and not the speech. We do
 335 not expect the AWGN interference to be a part of the signal model created by
 336 MP until much later iterations.

337 Fig. 10 shows $|S_{i_q}(t, M)|$ (28) aligned with the original signal for four values
 338 of M using the sparse approximations of the clean and distorted signals. We plot
 339 at the rear of these figures the localized magnitude time-domain correlation of
 340 the windowed and normalized signal with the query \mathbf{x}_q . In Fig. 10(a), using the
 341 clean \mathbf{x}_q , we clearly see its position even when using a single atom pair for each
 342 100 ms partition of the time-domain. We see the same behavior in Fig. 10(b)–

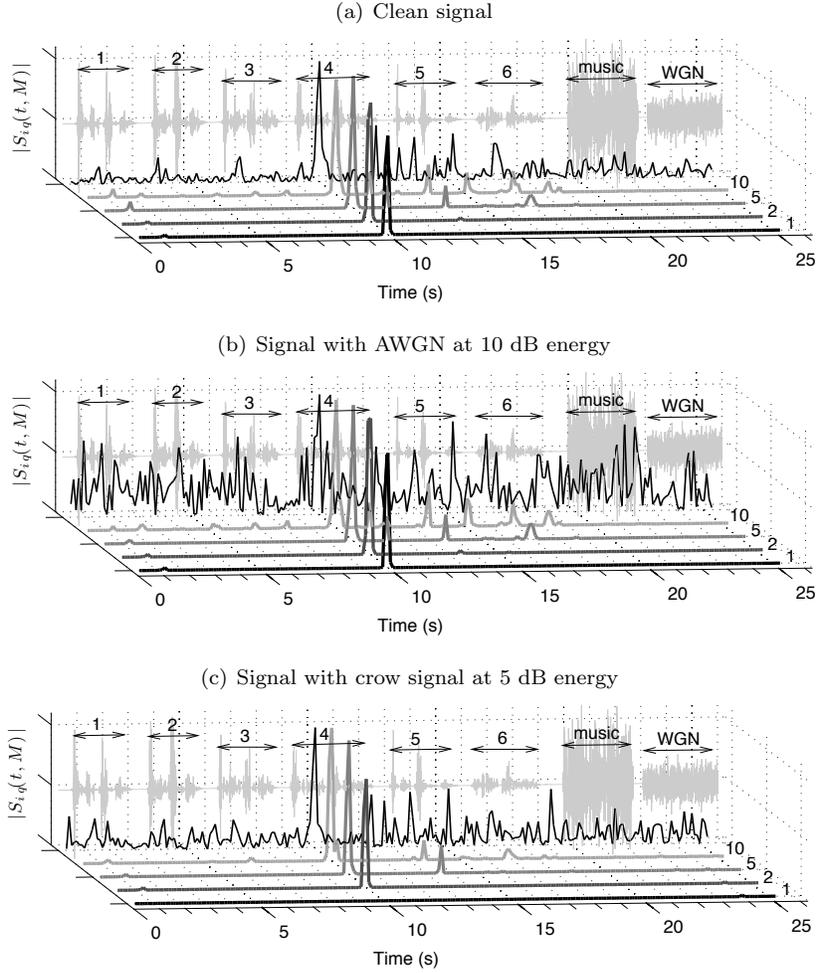


Figure 10: $|S_{iq}(M, t)|$ (28) as a function of time and the number of atoms M (labeled at right) considered from each representation for each localized sparse approximation. Localized magnitude correlation of each signal with query is shown by the thin black line in front of the gray time-domain signal at rear.

343 (c) for the two distorted signals, but in the case where the crow sound interferes
 344 we find the query for $M \geq 2$, or with at least three atom pair comparisons. The
 345 first atom of the decomposed query with the crow is modeling the crow and not
 346 the content of interest, and so we must increase the order of the model to find
 347 the location of \mathbf{x}_q . As we increase the number of pairs considered we also find

#	Clean Signal			Signal + WGN			Signal + Crow		
	t (s)	$ S_{iq} $	content	t (s)	$ S_{iq} $	content	t (s)	$ S_{iq} $	content
1	10.0	0.798	“cheese”	10.0	0.236	“cheese”	10.0	0.409	“cheese”
2	13.6	0.199	“cheese”	13.6	0.080	“cheese”	13.6	0.060	“cheese”
3	11.3	0.153	“-ives is-”	15.1	0.051	“delicious”	16.9	0.030	“cheese”
4	16.9	0.149	“cheese”	11.3	0.045	“-ives is-”	6.9	0.012	“cheese”
5	15.1	0.141	“delicious”	16.9	0.042	“cheese”	1.3	0.011	“cheese”
6	18.3	0.076	“delicious”	18.3	0.028	“delicious”	18.3	0.010	“delicious”
7	1.3	0.057	“cheese”	8.1	0.014	“delicious”	13.2	0.010	“cottage”
8	8.1	0.035	“delicious”	12.0	0.012	“-licious”	15.1	0.009	“delicious”
9	2.4	0.026	“delicious”	5.2	0.011	“delicious”	16.0	0.004	“cott-”
10	6.9	0.024	“cheese”	6.8	0.010	“cheese”	22.8	0.003	WGN

Table 2: Times, values and signal content for first 10 peaks in $|S_{iq}(t, 10)|$ ($P(10) = 55$) in Figs. 10(a)–(c). Highest-rated distances in each (bold) points to the origin of signal.

348 other segments that point in the same direction as \mathbf{x}_q . Table 2 gives the times
349 and content of the ten largest values in $|S_{iq}(t, 10)|$. For the clean and AWGN
350 distorted \mathbf{x}_q , “cheese” appears five of the six times it exists in the original signal.
351 Curiously, these same five instances are the five smallest distances when \mathbf{x}_q has
352 the crow interference.

353 We perform the same test as above but using a much longer speech signal
354 (about 21 minutes in length) excerpted from a book-on-CD, “The Old Man and
355 the Sea” by Ernest Hemingway, read aloud by a single person. From this signal
356 we create several queries \mathbf{x}_q , from words to sentences to an entire paragraph of
357 duration 35 s. We decompose each signal over the dictionary in Table 1 until the
358 global residual energy is 20 dB below the initial energy. The approximation of
359 the entire 21 m signal has 1,004,001 atoms selected from a dictionary containing
360 2,194,730,297 atoms.

361 One \mathbf{x}_q we extract from the signal is the spoken phrase, “the old man said”
362 (861 ms in length). This phrase appears 26 times in the long excerpt. We
363 evaluate $|S_{iq}(t, M)|$ (28) every 116 ms, and find the time at which \mathbf{x}_q originally
364 appears using only $M = 1$ atom pair comparisons for each time partition. The

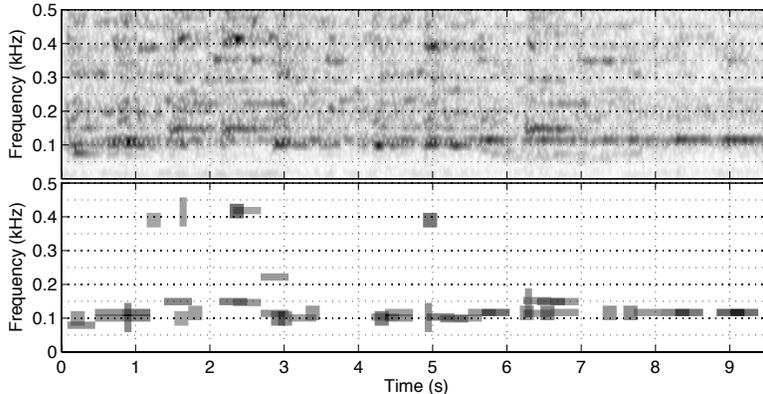


Figure 11: Polyphonic orchestral query: sonogram (top) and time-frequency tiles (bottom) of 50-order sparse approximation.

365 next highest ranked positions have values of 75% and 67% that of the largest
 366 $|S_{iq}(t, 1)|$. When $M = 50$, the values of the second and third largest values
 367 $|S_{iq}(t, 50)|$ drop to 62% and 61% that of the largest value. In the top 30 ranked
 368 subsequences for $M = 5$ we find only one of the other 25 appearances of “the old
 369 man said” (rank 26); but we also find “the old man agreed” (rank 11), and “the
 370 old man carried” (rank 16). All other results have minimal content similarity to
 371 the signal, but have time-frequency overlap in parts of the atoms of each model.

372 We perform the same test with a sentence extracted from the signal, “They
 373 were as old as erosions in a fishless desert” (2.87 s), which only appears once. No
 374 matter the $M = [1, 50]$ we use, the origin of the excerpt remains at a rank of 6
 375 with a value $|S_{iq}(t, 50)|$ at 67.5% that of the highest rank subsequence. We find
 376 that if we shift the time partition forward by 11.6 ms its ranking jumps to first,
 377 with the second ranked subsequence at 73%. We observe a similar effect for a
 378 query consisting of an entire paragraph (35 s). We find its origin by comparing
 379 $M = 2$ or more atoms from each model using a time partition of 116 ms. This
 380 result, however, disappears when we evaluate $|S_{iq}(t, M)|$ using a coarser time
 381 partition of 250 ms.

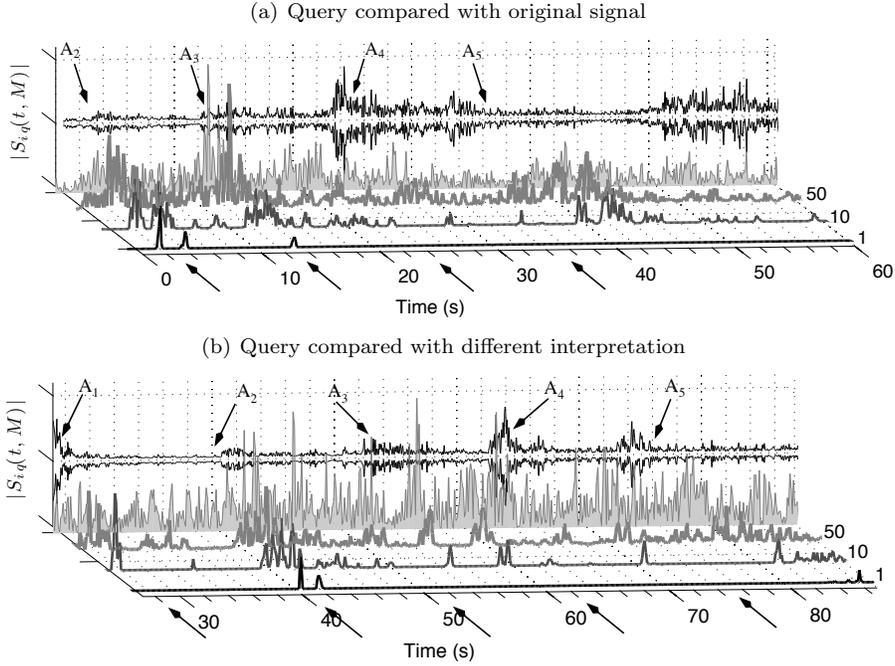


Figure 12: $|S_{iq}(t, M)|$ (28) for three values of M for the query and two different signals. Arrows mark the appearances of the ‘A’ theme, and their appearance number. Magnitude correlation of query with localized and normalized signal is shown by the solid gray area in front of the black time-domain signal at rear.

382 4.3. Experiment 3: Comparing Music Signals

383 While the previous experiment deals with single-channel speech signals, in
 384 this experiment we make comparisons between polyphonic musical signals ex-
 385cerpted from a commercial recording of the fourth movement of *Symphonie*
 386*Fantastique* by H. Berlioz. For the query, we use a 10.31 s segment of the third
 387appearance of the ‘A’ theme of the movement (bars 33 – 39, located around 13
 388– 22 s in Fig. 4(b)). Figure 11 shows the sonogram and time-frequency tiles of
 389the model of \mathbf{x}_q using the 50 atoms with the largest magnitude weights selected
 390from the 8xMDCT dictionary [31]. We add no interfering signals as we do in
 391the previous experiment.

392 Fig. 12(a) shows $|S_{iq}(t, M)|$ (28) over the first minute of the original signal,
 393 for three values of M , including $M = 50$, the time-frequency representation of

394 which is shown at bottom of Fig. 11. For $|S_{i_q}(t, 50)|$ we can see a strong spike
 395 located around 13 s corresponding with the query, but we also see spikes at
 396 about 2 s and around 43 s. The former set of spikes correspond with the second
 397 appearance of the ‘A’ theme, when only low bowed strings are playing the theme
 398 in G-minor. This is quite similar to the instrumentation of the query: low bowed
 399 strings and a legato bassoon in counterpoint in Eb-major. The latter set of spikes
 400 is around the end of the fifth appearance of the theme, which is played in G-
 401 minor on low pizzicato strings with a staccato bassoon. For $M = 10$, we see a
 402 conspicuous spike at the time of the fifth appearance around 34 s, as well as of
 403 the fourth appearance around 24 s, where the theme is played in Eb-major like
 404 the query. Finally, we test how the sparse approximation of this query compares
 405 with subsequences from a different recording of this movement, which is also in
 406 a different tempo. Figure 12(b) shows $|S_{i_q}(t, M)|$ (28) for three different values
 407 of M . We see high similarity with the first and second appearances of the main
 408 theme, but not the third, which is what the query contains.

409 4.4. Discussion

410 There is no reason to believe that a robust and accurate speech or melody
 411 recognition system can be created by comparing only the first few elements
 412 of greedy decompositions in time-frequency dictionaries. What appears to be
 413 occurring for the short signals, both the “cheese” and “the old man said,” is
 414 that the first few elements of their sparse and atomic decomposition create a
 415 prosodic representation that is comparable to others at the atomic level. For
 416 the longer signals, such as sentences, paragraphs, and orchestral theme, a few
 417 atoms cannot adequately embody the prosody, but we still see that by only
 418 making a few comparisons we are able to locate the excerpted signal — as long
 419 as the time partition is fine enough. This is due to the atoms of the models
 420 acting in some sense as a time-frequency fingerprint, an example of which is in
 421 Fig. 11. Through the cosine distance, the relative time and frequency locations
 422 of the atoms in the query and subsequence are being compared. Subsequences
 423 that share atoms in similar configurations will be gauged closer to \mathbf{x}_q than those

424 that do not.

425 By using the cosine distance it is not unexpected that (28) will be extremely
426 sensitive to a partitioning of the time-domain. This comes directly from the
427 definition of the time-localized Gramian (23), as well as the use of a dictionary
428 that is not translation invariant. There is no need to partition the time axis
429 when using a parameterized dictionary if we assume that some of the atoms
430 in the model of \mathbf{x}_q will have parameters that are nearly the same as some of
431 those in the relevant localized sparse representations. In such a scenario, we can
432 search a sparse representation for the times at which atoms exist that are similar
433 in scale and modulation frequency to those modeling \mathbf{x}_q . Then we can limit our
434 search to those particular times without considering any uniform and arbitrary
435 partition of the time-domain. With non-linear greedy decomposition methods
436 such as MP and time-variant dictionaries, however, such an assumption cannot
437 be guaranteed; but its limits are not yet well-known.

438 5. Conclusion

439 In this paper, we have extended and investigated the applicability of a
440 method of recursive nearest neighbor search [14, 15] for comparing audio signals
441 using pairwise comparisons of model elements in a sparse domain. The multi-
442 scale descriptions offered by sparse approximation over time-frequency dictio-
443 naries are especially attractive for such tasks because they provide flexibility in
444 making comparisons between data, not to mention a capacity to deal with noisy
445 signals. After extending this method to the task of comparing subsequences
446 of audio signals, we find that the strongest bound known for the remainder is
447 too weak to quickly and efficiently reduce the search space. Our experiments
448 show, however, that by comparing elements of sparse models we can judge
449 with relatively few comparisons whether signals share the same time-frequency
450 structures, and to what degrees, although this can be quite sensitive to the
451 time-domain partitioning. We also see that we can approach such comparisons
452 hierarchically, starting from the most energetic content to the least, or starting

453 from the longest scale phenomenon to the shortest.

454 We are continuing this research in multiple directions. First, since we know
 455 that the inner product matrix $\mathbf{G}_{iq}(t)$ (23) will be very sparse for all t in time-
 456 frequency dictionaries, this motivates designing a tighter bound based on a
 457 Laplacian distribution of elements in $\mathbf{G}_{iq}(t)$ with a large probability mass ex-
 458 actly at zero. This bound would be much more realistic than that provided by
 459 assuming the elements of the Gramian are distributed uniform (16). Another
 460 part of the problem is of course that the sums in (9) and (28) are not such that
 461 at step M the $P(M)$ largest magnitude values of $\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}$ are actually being
 462 summed. By our assumption in (4), we know that the decay of the magnitudes
 463 of the elements in \mathbf{A}_{iq} will be quickest in diagonal directions, but dependent
 464 upon the element position in the matrix. These diagonal directions are simply
 465 given by

$$\begin{bmatrix} \partial/\partial\gamma_i \\ \partial/\partial\gamma_q \end{bmatrix} m^{-\gamma_i} l^{-\gamma_q} = -m^{-\gamma_i} l^{-\gamma_q} \begin{bmatrix} \gamma_i/m \\ \gamma_q/l \end{bmatrix} \quad (29)$$

466 where we now recognize that the weights of two different representations can
 467 decay at different rates. With this, we can make an ordered set of index pairs
 468 by

$$\Lambda = \{(m, l)_\lambda : |[\mathbf{A}_{iq}]_\lambda| \geq |[\mathbf{A}_{iq}]_{\lambda+1}|\}_{\lambda=1,2,\dots,n_i n_q} \quad (30)$$

469 and define a recursive sum for $1 < m \leq n_i n_q$

$$S_{iq}(m) \triangleq S_{iq}(m-1) + [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{\Lambda_m} \quad (31)$$

470 setting $S_{iq}(1) = [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{11}$. We do not yet know the extent to which this
 471 approach can ameliorate the problems with the non-discriminating bound (16),
 472 as we have yet to design an efficient way to generate a satisfactory Λ , and
 473 estimate the bounds of the corresponding remainder — whether it is like that
 474 in (16), or another that uses the fact that $\mathbf{G}_{iq}(t)$ will be very sparse, even when
 475 $\mathbf{x}_q = \mathbf{y}_i$. We think that using a stronger bound and this indexing order will
 476 significantly reduce the number of pairwise comparisons that must be made

477 before determining a subsequence is not close enough with respect to the cosine
478 distance. Furthermore, we can make the elements of \mathbf{A}_{iq} decay faster, and thus
479 increase γ , by using other sparse approximation approaches, such as OMP [28,
480 23] or CMP [36]. And we cannot forget the implications of choosing a particular
481 dictionary. In this work we have used two different parametric dictionaries, one
482 of which is designed for audio signal coding [31]. Another interesting research
483 direction is to use dictionaries better suited for content description than coding,
484 such as content-adapted dictionaries [20, 2, 19].

485 Finally, and specifically with regards to the specific problem of similar-
486 ity search in audio signals, the cosine distance between time-domain samples
487 makes little sense because it is too sensitive to signal waveforms whereas hu-
488 man perception is not. Instead, many other possibilities exist for comparing
489 sparse approximation, such as comparing low-level histograms of atom param-
490 eters [7, 34]; comparing mid-level structures such as harmonics [9, 38, 8]; and
491 comparing high-level patterns of short atoms representing rhythm [32]. There
492 also exists the Matching Pursuit Dissimilarity Measure [25], where the atoms of
493 one sparse model are used to decompose another signal, and vice versa to see
494 how well they model each other. We are exploring these various possibilities
495 with regards to gauging more generally similarity in audio signals at multiple
496 levels of specificity within a sparse domain.

497 **6. Acknowledgments**

498 B. L. Sturm performed part of this research as a Chateaubriand Post-
499 doctoral Fellow (N. 634146B) at the Institut Jean Le Rond d’Alembert, Équipe
500 Lutheries, Acoustique, Musique, Université Pierre et Marie Curie, Paris 6,
501 France; as well as at the Department of Architecture, Design and Media Tech-
502 nology, at Aalborg University Copenhagen, Denmark. L. Daudet acknowledges
503 partial support from the French Agence Nationale de la Recherche under con-
504 tract ANR-06-JCJC-0027-01 DESAM. The authors thank the anonymous re-
505 viewers for their very detailed and helpful comments.

506 **Appendix A. Proof of Remainder Bounds**

507 To show (14), we can bound $R(M)$ loosely by assuming the worst case sce-
 508 nario of $[\mathbf{G}_{iq}]_{ml} = 1$ for all its elements. Knowing that $R(M)$ is the sum of the
 509 elements of the matrix $\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}$ except for the first $P(M)$ values, and assuming
 510 (4), we can say

$$C^{-2}R(M) \leq \sum_{m=M+1}^n \sum_{l=1}^m [l(m-l+1)]^{-\gamma} + \sum_{m=1}^{n-1} \sum_{l=m+1}^n [l(n-m+1)]^{-\gamma} \\ = \|\mathbf{c}_M^\gamma\|_1 + \|\mathbf{d}^\gamma\|_1. \quad (\text{A.1})$$

511 where \mathbf{c}_M^γ and \mathbf{d}^γ are defined in (17) and (18). This worst case scenario is not
 512 possible using MP because of its update rule (1).

513 We can find the tighter bound in (15) by assuming the distribution of
 514 signs of the elements of \mathbf{G}_{iq} is Bernoulli equiprobable, i.e., $P\{[\mathbf{G}_{iq}]_{ml} = 1\} =$
 515 $P\{[\mathbf{G}_{iq}]_{ml} = -1\} = 0.5$. Thus, defining a random variable $b_i : \mathbb{R} \mapsto \{-1, 1\}$,
 516 and its probability mass function $f_B(b_i) = 0.5\delta(b_i + 1) + 0.5\delta(b_i - 1)$ using the
 517 Dirac function, $\delta(x)$, we create a random vector \mathbf{b} with $n^2 - P(M)$ elements
 518 independently drawn from this distribution. Placing this into the double sums
 519 of (A.1) provides the bound

$$C^{-2}R(M) \leq \left| \mathbf{b}^T \begin{bmatrix} \mathbf{c}_M^\gamma \\ \mathbf{d}^\gamma \end{bmatrix} \right| \leq \|\mathbf{c}_M^\gamma\|_1 + \|\mathbf{d}^\gamma\|_1. \quad (\text{A.2})$$

520 This weighted Rademacher sequence has the property that [14]

$$P\{|\mathbf{b}^T \mathbf{s}| > R\} \leq 2 \exp(-R^2/2\|\mathbf{s}\|_2^2), R > 0 \quad (\text{A.3})$$

521 which becomes $P\{|\mathbf{b}^T \mathbf{s}| \leq R\} \geq \max\{0, 1 - 2 \exp(-R^2/2\|\mathbf{s}\|_2^2)\}$ by the ax-
 522 ioms of probability. With this we can find an R such that $P\{|\mathbf{b}^T \mathbf{s}| \leq R\}$ will
 523 be greater than or equal to some probability $0 \leq p \leq 1$, i.e.,

$$R(p) = (\|\mathbf{c}_M^\gamma\|_2^2 + \|\mathbf{d}^\gamma\|_2^2)^{1/2} \left[2 \ln \frac{2}{1-p} \right]^{1/2}. \quad (\text{A.4})$$

524 This value can be minimized by choosing $p = 0$, for which we arrive at the
 525 residual upper bound in (15). Note that even though we have set $p = 0$, we still

526 have an unrealistically loose bound by the impossibility of MP of choosing two
 527 sets of atoms for which all entries of their Gramian \mathbf{G}_{i_q} are in $\{-1, 1\}$.

528 Finally, to show (16), we can model the elements of the Gramian as random
 529 variables, $u_i : \mathbb{R} \mapsto [-1, 1]$, independently and identically distributed uniformly

$$f_U(u_i) = \begin{cases} 0.5, & -1 \leq u_i \leq 1 \\ 0, & \text{else.} \end{cases} \quad (\text{A.5})$$

530 Substituting this into (14) gives a weighted sum of random variables satisfying

$$C^{-2}R(M) \leq \left| \mathbf{u}^T \begin{bmatrix} \mathbf{c}_M^\gamma \\ \mathbf{d}^\gamma \end{bmatrix} \right| \leq \|\mathbf{c}_M^\gamma\|_1 + \|\mathbf{d}^\gamma\|_1. \quad (\text{A.6})$$

531 where \mathbf{u} is the random vector. For large M , this sum has the asymptotic
 532 property [14, 15]:

$$P\{|\mathbf{u}^T \mathbf{s}| < R\} = \text{Erf} \sqrt{\frac{3R^2}{2\|\mathbf{s}\|_2^2}}. \quad (\text{A.7})$$

533 Setting this equal to $0 \leq p \leq 1$ and solving for R produces the upper bound
 534 (16). We can reach the upper bound (15) if we set $p = 0.9586$, but note that
 535 (16) can be made zero. This bound can still be extremely loose because the
 536 Gramian of two models in time-frequency dictionaries will be highly sparse.

537 Computing the ℓ_2 -norm in these expressions, however, leads to evaluating
 538 the double sums

$$\|\mathbf{c}_M^\gamma\|^2 = \sum_{m=M+1}^n \sum_{l=1}^m \frac{1}{[l(m-l+1)]^{2\gamma}} \quad (\text{A.8})$$

$$\|\mathbf{d}^\gamma\|^2 = \sum_{m=1}^{n-1} \sum_{l=m+1}^n \frac{1}{[l(n-m+1)]^{2\gamma}} \quad (\text{A.9})$$

539 which can be prohibitive for large n . The dimensionality of \mathbf{c}_M^γ is $n(n+1)/2 -$
 540 $P(M)$, and of \mathbf{d}^γ is $n(n-1)/2$. We approximate these values in the following
 541 way for $\gamma = 0.5$, using the partial sum of the harmonic series

$$\sum_{m=1}^n \frac{1}{m} = \ln n + \gamma_E + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} + \mathcal{O}(n^{-6}) \quad (\text{A.10})$$

542 where $\gamma_E \approx 0.5772$ is the Euler-Mascheroni constant. To find $\|\mathbf{d}^{0.5}\|^2$

$$\begin{aligned}
\|\mathbf{d}^{0.5}\|^2 &= \sum_{m=1}^{n-1} \sum_{l=m+1}^n \frac{1}{l(n-m+1)} \\
&= \sum_{m=1}^{n-1} \frac{1}{n-m+1} \left[\sum_{l=1}^n \frac{1}{l} - \sum_{l=1}^m \frac{1}{l} \right] \\
&\approx \sum_{m=1}^{n-1} \frac{1}{n-m+1} \left[\ln n/m - \frac{n-m}{2nm} + \frac{n^2-m^2}{12n^2m^2} \right]. \tag{A.11}
\end{aligned}$$

543 To find $\|\mathbf{c}_M^{0.5}\|^2$ we first use partial fractions and then the partial sum of the
544 harmonic series:

$$\begin{aligned}
\|\mathbf{c}_M^{0.5}\|^2 &= \sum_{m=M+1}^n \sum_{l=1}^m \frac{1}{l(m-l+1)} \\
&= \sum_{m=M+1}^n \frac{1}{m+1} \sum_{l=1}^m \frac{1}{l} + \frac{1}{m-l+1} \\
&\approx \sum_{m=M+1}^n \frac{1}{m+1} \left[\ln m + \gamma_E + \frac{1}{2m} - \frac{1}{12m^2} + \sum_{l=1}^m \frac{1}{l} \right] \\
&\approx 2 \sum_{m=M+1}^n \frac{1}{m+1} \left(\ln m + \gamma_E + \frac{1}{2m} - \frac{1}{12m^2} \right). \tag{A.12}
\end{aligned}$$

545 With these expressions we can avoid double sums in calculating the bounds.

546 Appendix B. Estimating the Compressibility Parameters

547 The compressibility parameters (C, γ) must be estimated for the set of
548 weights in $\hat{\mathcal{Y}}$ (??), as well as those of \mathbf{x}_q . Since by (4) the parameters (C, γ)
549 bound from above the decay of all the ordered weights, only the largest mag-
550 nitude weights matter for their estimation. Thus, we define a vector, \mathbf{a} , of the
551 largest n magnitude weights from each row in the set $\{\{\mathbf{a}_i(n_i)\}_{i \in \mathcal{I}}, \mathbf{a}_q(n_q)\}$,
552 which is equivalent to taking the largest weights at each approximation order.
553 Good compressibility parameters can be given by

$$\min_{C, \gamma} \|C\mathbf{z}^\gamma - \mathbf{a}\|^2 + \lambda C \text{ subject to } C\mathbf{z}^\gamma \succeq \mathbf{a} \tag{B.1}$$

554 where we define $\mathbf{z}^\gamma \triangleq [1, 1/2^\gamma, \dots, 1/n^\gamma]^T$, and add a multiple of C in order to
555 keep it from getting too large since the bounds (14)–(16) are all proportional to

556 it. The constraint is added to ensure all elements of the difference $C\mathbf{z}^\gamma - \mathbf{a}_i$ are
 557 positive such that (4) is true.

558 To remove the γ component from the exponent, and since all of the elements
 559 of \mathbf{z} and \mathbf{a} are positive and non-zero, we can instead solve the problem

$$\begin{aligned} & \min_{C, \gamma} \|\ln C\mathbf{1} + \gamma \ln \mathbf{z} - \ln \mathbf{a}\|^2 + \lambda \ln C \\ & = \min_{C, \gamma} \left[(\ln C)^2 n + \gamma^2 \|\ln \mathbf{z}\|^2 + \|\ln \mathbf{a}\|^2 + \lambda \ln C \right. \\ & \quad \left. + 2\gamma (\ln \mathbf{z})^T (\ln C\mathbf{1} - \ln \mathbf{a}) - 2 \ln C (\ln \mathbf{a})^T \mathbf{1} \right] \quad (\text{B.2}) \end{aligned}$$

560 subject to the constraint $C\mathbf{z}^\gamma \succeq \mathbf{a}$. Taking the partial derivative of this with
 561 respect to γ and C , we find

$$\gamma_o = \frac{(\ln \mathbf{z})^T (\ln \mathbf{a} - \ln C\mathbf{1})}{\|\ln \mathbf{z}\|^2} \quad (\text{B.3})$$

$$C_o = \exp \left[\lambda + \frac{1}{n} \sum_{i=1}^n [\ln \mathbf{a} - \gamma \ln \mathbf{z}]_i \right]. \quad (\text{B.4})$$

562 Starting with some initial value of C then, we use the following iterative method

- 563 1. solve for γ given a C in (B.3);
- 564 2. find the new C in (B.4) using this γ ;
- 565 3. set $C' = \exp [\max(\ln \mathbf{a} - \gamma_o \ln \mathbf{z})]$ and evaluate the error $\|C'\mathbf{z}^\gamma - \mathbf{a}\|^2$;
- 566 4. repeat until the error begins to increase.

567 The factor λ in effect controls the step size for convergence. A typical value
 568 we use is $\lambda = \pm 0.03$ based on experiments (the sign of which depends on if the
 569 objective function decreases with decreasing C).

570 [1] Agrawal, R., Faloutsos, C., Swami, A., Oct. 1993. Efficient similarity search
 571 in sequence databases. In: Proc. Int. Conf. Foundations Data Org. Algo.
 572 Chicago, IL, pp. 69–84.

573 [2] Aharon, M., Elad, M., Bruckstein, A., Nov 2006. K-SVD: An algorithm
 574 for designing of overcomplete dictionaries for sparse representation. IEEE
 575 Trans. Signal Process. 54 (11), 4311–4322.

- 576 [3] Casey, M., Rhodes, C., Slaney, M., July 2008. Analysis of minimum dis-
577 tances in high-dimensional musical spaces. *IEEE Trans. Audio, Speech,*
578 *Lang. Process.* 16 (5), 1015–1028.
- 579 [4] Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M.,
580 Apr. 2008. Content-based music information retrieval: Current directions
581 and future challenges. *Proc. IEEE* 96 (4), 668–696.
- 582 [5] Chang, K., Jang, J.-S. R., Iliopoulos, C. S., Aug. 2010. Music genre clas-
583 sification via compressive sampling. In: *Proc. Int. Soc. Music Information*
584 *Retrieval.* Amsterdam, The Netherlands, pp. 387–392.
- 585 [6] Chen, S. S., Donoho, D. L., Saunders, M. A., Aug. 1998. Atomic decompo-
586 sition by basis pursuit. *SIAM J. Sci. Comput.* 20 (1), 33–61.
- 587 [7] Chu, S., Narayanan, S., Kuo, C.-C. J., Aug. 2009. Environmental
588 sound recognition with time-frequency audio features. *IEEE Trans. Audio,*
589 *Speech, Lang. Process.* 17 (6), 1142–1158.
- 590 [8] Cotton, C., Ellis, D. P. W., Oct. 2009. Finding similar acoustic events using
591 matching pursuit and locality-sensitive hashing. In: *Proc. IEEE Workshop*
592 *App. Signal Process. Audio and Acoustics.* Mohonk, NY, pp. 125–128.
- 593 [9] Daudet, L., Sep. 2006. Sparse and structured decompositions of signals
594 with the molecular matching pursuit. *IEEE Trans. Audio, Speech, Lang.*
595 *Process.* 14 (5), 1808–1816.
- 596 [10] Ellis, D. P. W., Poliner, G. E., Apr. 2007. Identifying ‘cover songs’ with
597 chroma features and dynamic programming beat tracking. In: *Proc. Int.*
598 *Conf. Acoustics, Speech, Signal Process.* Honolulu, Hawaii, pp. 1429–1432.
- 599 [11] Faloutsos, C., Ranganathan, M., Manolopoulos, Y., 1994. Fast subsequence
600 matching in time-series databases. In: *Proc. ACM SIGMOD Int. Conf.*
601 *Mgmt. Data.* Minneapolis, MN, pp. 419–429.

- 602 [12] Gemmeke, J., ten Bosch, L., L.Boves, Cranen, B., Aug. 2009. Using sparse
603 representations for exemplar based continuous digit recognition. In: Proc.
604 EUSIPCO. Glasgow, Scotland, pp. 1755–1759.
- 605 [13] Haitsma, J., Kalker, T., June 2003. A highly robust audio fingerprinting
606 system with an efficient search strategy. *J. New Music Research* 32 (2),
607 211–221.
- 608 [14] Jost, P., June 2007. Algorithmic aspects of sparse approximations. Ph.D.
609 thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
- 610 [15] Jost, P., Vandergheynst, P., Aug. 2008. On finding approximate nearest
611 neighbours in a set of compressible signals. In: Proc. European Signal
612 Process. Conf. Lausanne, Switzerland, pp. 1–5.
- 613 [16] Kimura, A., Kashino, K., Kurozumi, T., Murase, H., Feb. 2008. A quick
614 search method for audio signals based on piecewise linear representation
615 of feature trajectories. *IEEE Trans. Audio, Speech, Lang. Process.* 16 (2),
616 396–407.
- 617 [17] Krstulovic, S., Gribonval, R., Apr. 2006. MPTK: Matching pursuit made
618 tractable. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Vol. 3.
619 Toulouse, France, pp. 496–499.
- 620 [18] Kurth, F., Müller, M., Feb. 2008. Efficient index-based audio matching.
621 *IEEE Trans. Audio, Speech, Lang. Process.* 16 (2), 382–395.
- 622 [19] Leveau, P., Vincent, E., Richard, G., Daudet, L., Jan. 2008. Instrument-
623 specific harmonic atoms for mid-level music representation. *IEEE Trans.*
624 *Audio, Speech, Lang. Process.* 16 (1), 116–128.
- 625 [20] Lewicki, M. S., Sejnowski, T. J., Feb. 2000. Learning overcomplete repre-
626 sentations. *Neural Computation* 12, 337–365.
- 627 [21] Li, C.-S., Yu, P. S., Castelli, V., Feb. 1996. Hierarchyscan: A hierarchical
628 similarity search algorithm for databases of long sequences. In: Proc. Int.
629 Conf. Data Eng. New Orleans, LA, pp. 546–553.

- 630 [22] Lyon, R. F., Rehn, M., Bengio, S., Walters, T. C., Chechik, G., Sep. 2010.
631 Sound retrieval and ranking using sparse auditory representations. *Neural*
632 *Computation* 22 (9), 2390–2416.
- 633 [23] Mailhé, B., Gribonval, R., Vandergheynst, P., Bimbot, F., 2011. Fast or-
634 thogonal sparse approximation algorithms over local dictionaries. *Signal*
635 *Process.* (accepted).
- 636 [24] Mallat, S., 2009. *A Wavelet Tour of Signal Processing: The Sparse Way*,
637 3rd Edition. Academic Press, Elsevier, Amsterdam.
- 638 [25] Mazhar, R., Gader, P. D., Wilson, J. N., Oct. 2009. Matching pursuits
639 dissimilarity measure for shape-based comparison and classification of high-
640 dimensional data. *IEEE Trans. Fuzzy Syst.* 17 (5), 1175–1188.
- 641 [26] Müller, M., Kurth, F., Clausen, M., Oct. 2005. Chroma-based statistical
642 audio features for audio matching. In: *IEEE Workshop on Applications of*
643 *Signal Processing to Audio and Acoustics*. New Paltz, NY, pp. 275–278.
- 644 [27] Panagakis, Y., Kotropoulos, C., Arce, G. R., Aug. 2009. Music genre clas-
645 sification via sparse representations of auditory temporal modulations. In:
646 *Proc. European Signal Process. Conf.* Glasgow, Scotland, pp. 1–5.
- 647 [28] Pati, Y., Rezaiifar, R., Krishnaprasad, P., Nov. 1993. Orthogonal matching
648 pursuit: Recursive function approximation with applications to wavelet
649 decomposition. In: *Proc. Asilomar Conf. Signals, Syst., Comput.* Vol. 1.
650 Pacific Grove, CA, pp. 40–44.
- 651 [29] Pham, T. V., Smeulders, A., Apr. 2006. Sparse representation for coarse
652 and fine object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 28 (4),
653 555–567.
- 654 [30] Rafiei, D., Mendelzon, A., Nov. 1998. Efficient retrieval of similar time
655 sequences using DFT. In: *Proc. Int. Conf. Found. Data Org.* Kobe, Japan,
656 pp. 249–257.

- 657 [31] Ravelli, E., Richard, G., Daudet, L., Nov. 2008. Union of MDCT bases for
658 audio coding. *IEEE Trans. Audio, Speech, Lang. Proc.* 16 (8), 1361–1372.
- 659 [32] Ravelli, E., Richard, G., Daudet, L., Mar. 2010. Audio signal representa-
660 tions for indexing in the transform domain. *IEEE Trans. Audio, Speech,
661 Lang. Process.* 18 (3), 434–446.
- 662 [33] Rebollo-Neira, L., Lowe, D., Apr. 2002. Optimized orthogonal matching
663 pursuit approach. *IEEE Signal Process. Lett.* 9 (4), 137–140.
- 664 [34] Scholler, S., Purwins, H., Oct. 2010. Sparse coding for drum sound classifi-
665 cation and its use as a similarity measure. In: *Proc. Int. Workshop Machine
666 Learning Music ACM Multimedia*. Firenze, Italy.
- 667 [35] Serrà, J., Gómez, E., Herrera, P., Serra, X., Aug. 2008. Chroma binary
668 similarity and local alignment applied to cover song identification. *IEEE
669 Trans. Audio, Speech, Lang. Process.* 16, 1138–1151.
- 670 [36] Sturm, B. L., Christensen, M., Nov. 2010. Cyclic matching pursuit with
671 multiscale time-frequency dictionaries. In: *Proc. Asilomar Conf. Signals,
672 Systems, and Computers*. Pacific Grove, CA.
- 673 [37] Sturm, B. L., Shynk, J. J., Mar. 2010. Sparse approximation and the pur-
674 suit of meaningful signal models with interference adaptation. *IEEE Trans.
675 Audio, Speech, Lang. Process.* 18 (3), 461–472.
- 676 [38] Sturm, B. L., Shynk, J. J., McLeran, A., Roads, C., Daudet, L., June 2008.
677 A comparison of molecular approaches for generating sparse and struc-
678 tured multiresolution representations of audio and music signals. In: *Proc.
679 Acoustics*. Paris, France, pp. 5775–5780.
- 680 [39] Tzanetakis, G., Cook, P., July 2002. Musical genre classification of audio
681 signals. *IEEE Trans. Speech Audio Process.* 10 (5), 293–302.
- 682 [40] Umaphy, K., Krishnan, S., Jimaa, S., Apr. 2005. Multigroup classification
683 of audio signals using time-frequency parameters. *IEEE Trans. Multimedia
684* 7 (2), 308–315.

- 685 [41] Vincent, P., Bengio, Y., July 2002. Kernel matching pursuit. *Machine*
686 *Learning* 48 (1), 165–187.
- 687 [42] Wang, A., Oct. 2003. An industrial strength audio search algorithm. In:
688 *Proc. Int. Conf. Music Info. Retrieval*. Baltimore, Maryland, USA, pp. 1–4.
- 689 [43] Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T., Yan, S., June 2009.
690 *Sparse representation for computer vision and pattern recognition*. *Proc.*
691 *IEEE* 98 (6), 1031–1044.
- 692 [44] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., Ma, Y., Feb. 2009.
693 *Robust face recognition via sparse representation*. *IEEE Trans. Pattern*
694 *Anal. Machine Intell.* 31 (2), 210–227.