COMPARISON OF METHODS FOR SPARSE REPRESENTATION OF MUSICAL SIGNALS

Line Ørtoft Endelt and Anders la Cour-Harbo

Aalborg University
Department of Control Engineering
Frb. Vej 7C, 9220 Aalborg East, Denmark
{oertoft, alc}@control.aau.dk

ABSTRACT

Within the last few decades a number of new signal processing tools has appeared. These have mainly been compared using constructed signals, signals designed to show the advantage of a new method over already existing methods. In this paper we evaluate the methods Basis Pursuit, Minimum Fuel Neural Networks, Matching Pursuit, Best Orthogonal Basis, Alternating Projections and Methods of Frames on "real" signals. The methods are applied on a number of excerpts sampled from a small collection of music, and their ability to expres music signals in a sparse manner is evaluated. The sparseness is measured by a number of sparseness measures and results are shown on the ℓ^1 norm of the coefficients, using a dictionary containing a Dirac basis, a Discrete Cosine Transform, and a Wavelet Packet. Evaluated only on the sparseness Matching Pursuit is the best method, and it is also relatively fast.

1. INTRODUCTION

The results presented here are obtained as part of a research project on Automatic Classification of music. The idea is that by finding a sparse representation of music signals, i.e. a representation containing only a few significant elements, good features, that capture the nature of each particular piece of music, can be found. Many different methods for feature extraction from music or other sound signals exists, e.g. [1], [2] and [3]. In most of these methods representations are found by using the Fourier or Wavelet transforms, and by various kinds of filtering. Classification rates lie between 60 % for categorizing into 10 categories to about 90 % for classifying into 2-3 classes, but the tests are performed on samples of very different size and content, and cannot be compared directly.

The vectors corresponding to a Fourier (or Wavelet) transform have similar structure; the idee is that by representing a music signal in a redundant set of vectors, where the vectors have different structure, a more compact representation can be achieved, since music contains events of both short and long duration.

The music signals are considered to be elements in $\boldsymbol{R}^n$. A dictionary for $\boldsymbol{R}^n$ is a set of vectors that span $\boldsymbol{R}^n$ and the vectors in a dictionary are called atom. Note that there is no requirement on the size of the set, the smallest possible dictionary contain exactly n elements and is a basis of $\boldsymbol{R}^n$. A representation of a signal $\boldsymbol{b}$ in a dictionary $\boldsymbol{D}$, with m elements, is a vector $\boldsymbol{x} \in \boldsymbol{R}^m$, satisfying,

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{1}$$

where A is an $n \times m$ matrix having the vectors in $\boldsymbol{D}$ as its columns. The vector $\boldsymbol{x}$ contains the coefficients of the representation. When $\boldsymbol{D}$ contain more the n elements, this representation is not unique. This on one hand leads to added flexibility in choice of representation, on the other leads to more complexity in finding the representations.

2. METHODS

The results presented here are part of the results of a large test setup designed to evaluate signal representation methods, dictionaries and optimal signal length for sparse representation of music. Presenting all the results are to extensive for just one paper, but to justify the results the hole setup is described.

Some of the minimization methods are very time consuming, so only a limited number of songs and dictionaries are considered. Since storing all the representations found for all the excerpts requires too much storage capacity, a number of sparseness measures are used to evaluate the minimization methods. The rest of this section is a description of the setup, which minimization methods are applied, which dictionaries and sparseness measures and how the music is sampled.

2.1. Minimization Methods

Six different minimization methods for finding representations as in (1) are applied. Basis Pursuit [4], BP is an optimization method, which through interior point linear programming seeks the solution $\boldsymbol{x}$ to equation (1) having the smallest ℓ^1 norm.

Minimum Fuel Neural Networks [5, 6], MFNN also seek the solution with the smallest ℓ^1 norm, here the problem is rewritten into two coupled non-linear differential equations for iterating the solution.

Matching Pursuit [7], MP makes a full decomposition (an analysis) of the signal in the dictionary, in other word finds the inner products between the signal and the atoms, the atom corresponding to the highest inner product is chosen, and the signal is replaced by the original signal subtracted the projection of the signal in the direction of the chosen atom. The procedure is continued, using the new signal, until the residual contains less than 1% of the energy of the original signal.

Using Alternating Projections, AP the signal is decomposed in one of the bases of the dictionary (see sec. 2.2) a number of the atoms corresponding the highest inner products are chosen, the signal is replaced by the original signal subtracted the projections in the direction of these atoms. The procedure is continued going through all the bases until only a certain fraction of the energy is left in the signal.

Best Orthogonal Basis [8], BOB chooses the basis among all the bases in the dictionary corresponding to the coefficient vector having the smallest ℓ^1 norm (other minimization measures can be used). This method applies only to CP and WP Dictionaries (see sec. 2.2).

Method of Frames [9], MOF also known as the Moore Penrose Inverse or the generalized inverse, finds the representation corresponding to the coefficient vector with the smallest energy, i.e. having the smallest ℓ^2 norm.

All the representations are found using already existing Matlab functions (for references see [4] and [5]), which have been adjusted to this test setup.

2.2. Dictionaries

The dictionaries applied are concatenated of one, two or three of the following subdictionaries: Discrete Cosine Transform (DCT) over sampled by a factor two, the Krönecher basis (or Dirac basis, DIRAC), a Wavelet Packet (WP) generated using the coiflet wavelet with filter length 12 (the choice is made based on the considerations in [10]), and a cosine packet (CP) containing locally trigonometric cosine functions generated with a "sine bell". The way the subdictionaries are built, they contain a lot of orthonormal bases, which are essential for some of the minimization methods.

Five different dictionaries are applied 1: {DCT, DIRAC}, 2: {DCT, WP, DIRAC}, 3: {WP}, 4: {CP} and 5: {WP, CP, DIRAC}.

The four subdictionaries are supposed to describe different elements in a music signal, noise cannot be compressed, so the DIRAC basis is believed to describe the random noise in the signal. The DCT describes frequencies over the whole time interval, while the CP describes frequencies over local dyadic intervals of the signal. The WP is good at describing both rapid changes (short duration events) in the signal, which appears at e.g. a note onset, and long duration events. These assumptions are supported by the results in [11].

2.3. Music

The music samples considered originates from five different pieces of music. This is a very small part of the class of music signals, but the limitation is made due to time considerations, and the purpose of these preliminary tests is to choose the methods and dictionaries to focus on. The five music pieces are chosen from very different music genres, so it is possible to compare the different music pieces, and at least get an idea of whether it is possible to distinguish different classes of music by measuring the sparseness in different representations. The five music pieces are listed in table 1.

Table 1. The five pieces of music applied in the test.

Performer/composer	Title
Jean Michel Jarre	The Chronologie - Part 2
Joe Satriani	The Extremist (The Extremist)
Unknown	Jazz
Brahms	Ein Deutches Requiem - Part 2
Cher	The Power (Believe)

2.4. Length of music excerpts

Ten different excerpt lengths are applied, $2^7, 2^8, \ldots, 2^{16}$. The sampling rate is 44,1 kHz. For each length successive excerpts are sampled from the beginning of each song, there is no overlap between the excerpts of the same length. Between 20 and 160 excerpts are sampled of each length, most for short lengths. One excerpt cover between 2.9 ms and 1.49 sec, and the longest interval covered by successive excerpts is about 30 sec.

2.5. Sparseness Measures

A good measure of how sparse a representation is, is the ℓ^0 norm, since it counts the number of coefficients different from 0. But for most of the methods this is either n or n times the redundancy of the dictionary due to the nature of the methods. Therefore a number of other sparseness mea−

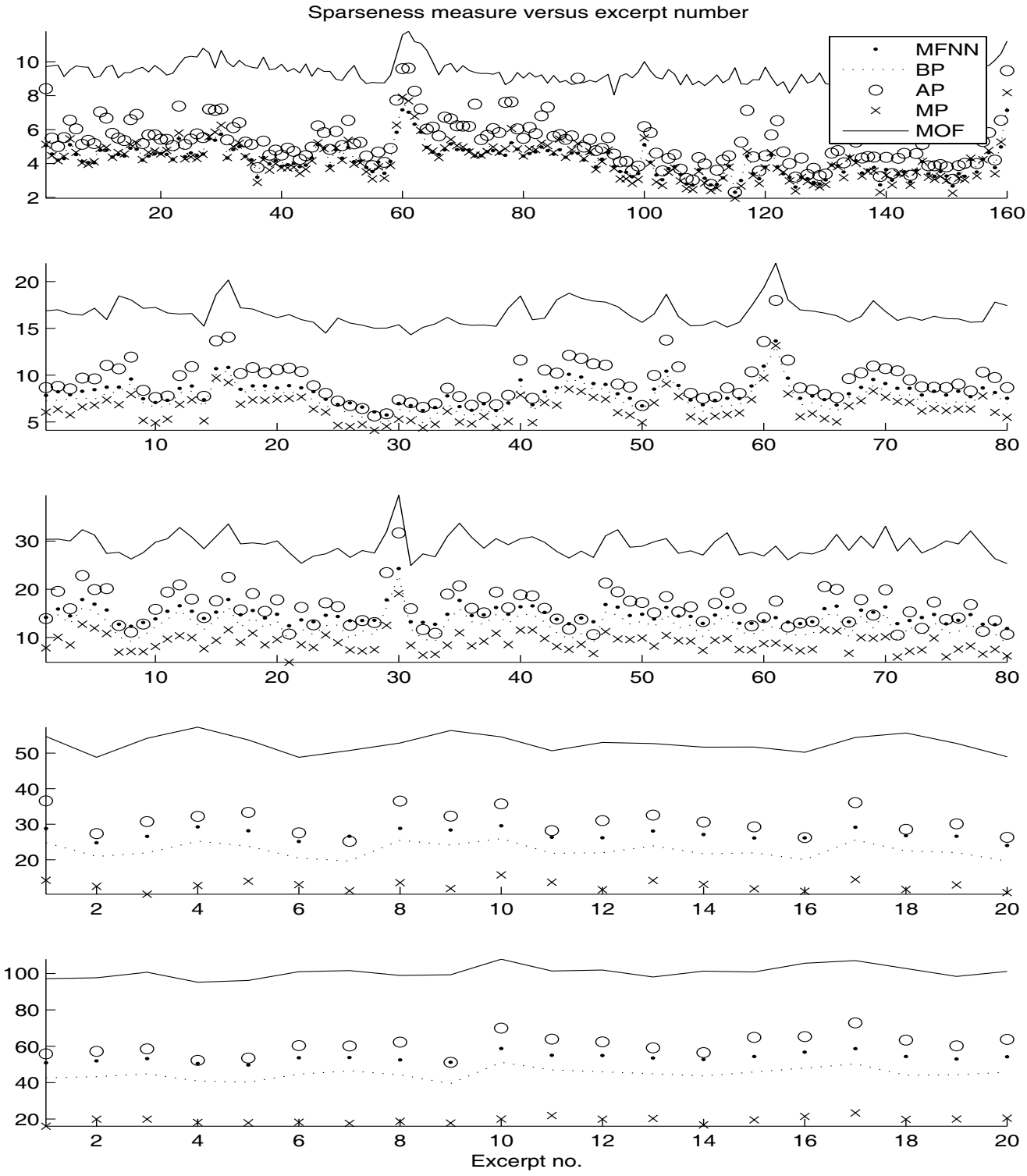

Fig. 1. The ℓ^1 norm of the representations found for Chers Believe for five different excerpt lengths, 256, 1024, 4096, 16384 and 65536, versus the excerpt no.

sures are applied $\ell^{0.5}$,$\ell^{0.6}$,$\ell^{0.7}$,$\ell^{0.8}$,$\ell^{0.9}$,ℓ^1, Shannon Entropy, Coifman-Wickerhauser entropy, Kurtosis, $\sum_i \log(1+x_i^2)$, $\sum_i \tanh(|x_i - \mathrm{median(x_i)}|)$, ℓ^2, ℓ^1/ℓ^2. The measures have been used within different areas of mathematics, but are all measures of the "concentration" of the energy in the coefficients.

3. RESULTS

The sparseness measures calculated for all the representations are stored in a six dimensional $6\times160\times10\times15\times5\times5$ array ([Method] $\times$ [excerpt no.] $\times$ [length no.] $\times$ [time, sparseness measure] $\times$ [Dictionary] $\times$ [music no.]).

A small part of the results are presented in Figure 1. Here the ℓ^1 norm of the representations found for the normed excerpts of length 256, 1024, 4096, 16384 and 65536, using dictionary no. 2 (DCT,WP,DIRAC). The excerpts cover 0.93, 1.86, 7.43, 7.43 and 29.7 seconds respectively, of the song "Believe" by Cher. All the results are very similar, so the relative placement of the data in Figure 1 do not change much if it was a different song or a different dictionary. Most of the sparseness measures give a similar picture, but some seems to favor methods like MP and AP giving a relative small part of coefficients different from 0.

For all length MOF gives the highest ℓ^1 norm, and except for the ℓ^2 norm this performs poorly for most sparseness measures. For short signals the ℓ^1 norm of the other four methods are at the same level, but for longer signal length, MP has a significantly smaller ℓ^1 norm. The three methods AP, MFNN and BP are at the same level with BP a little better than the others.

4. DISCUSSION

MP is expected to do better than AP, since there is more flexibility in the choice of atoms, it also takes longer time, since the analysis of the residual vector has to be performed more times, but still MP is relatively fast. The two methods MFNN and BP are both very time consuming and have about the same level of performance, the relative high ℓ^1 norm may be caused by the fact, that the methods are optimization methods leaving almost all coefficients different from 0, whereas MP chooses a number of atoms to represent the signal and the main part of coefficients are 0.

Another aspect in the judgment of the methods is the resolution. In [4] a number of constructed examples are shown, where BP performs significantly better than MP in resolving the signals. BP gives a good resolution of a music signal, and can separate a song into the beat and the main body or rhythm of the signal (an example can be found on http:/www.control.aau.dk/∼oertoft), but whether this is important when put into the statistical framework, which feature extraction necessarily has to be, only time will tell.

The computations are performed as distributed computations on a number of PCs, having very different CPU-power, so a comparison of the time consummation of the methods are not possible based on the data, but computations on the longest signals which takes about half and hour for MP can take about a day for BP (and MFNN). With the fast development in computer power, this may be an obstacle that can be overcome, and depends at the task at hand. The time requirements depend on whether the classification is to be performed in a few seconds, or more time can be allowed.

5. REFERENCES

[1] S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 619–625, September 2000.

[2] E.D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of Acoustical Society of America*, pp. 419–429, January 1998.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[4] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[5] A. la Cour-Harbo, "Application of the minimum fuel network to music signals," in *Proc.Int.Conf. on Acou., Speech and Signal Proc.*, 2004.

[6] Z. Wang, J. Cheung, Y. Xia, and J. Chen, "Minimum fuel neural network and their applications to overcomplete signal representation," *IEEE Trans. Circuit and Systems*, vol. 47, no. 8, pp. 1146–1159, August 2000.

[7] S. Mallet and Z. Zhang, "Matching pursuit in a time-frequency dictionary," *IEEE Transactions on Signal Processing*, pp. 3397–3415, 1993.

[8] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A K Peters, 1994.

[9] I. Daubechies, "Time-frequency localization operators: A geometric phase space approach," *IEEE Trans. Inform. Theory*, , no. 34, pp. 605–612, 1988.

[10] L. Ø. Endelt and A. la Cour-Harbo, "Wavelets for sparse representation of music," in *Proceedings of Wedelmusic2004*, 2004.

[11] L. Daudet, M. Sandler, and B. Torresani, "Audio representation on overcomplete sets," *Proceedings of 14th conf. on Digital Signal Processing.*, 2002.