Wavelets for Sparse Representation of Music

Endelt, Line Ørtoft; la Cour-Harbo, Anders

Published in:
Proceedings of WEDELMUSIC 2004

Publication date:
2004

Document Version
Publisher's PDF, also known as Version of record

Wavelets for Sparse Representation of Music

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

When using a discrete wavelet transform or a wavelet packet for obtaining a sparse representation of music signals the first question that arises is which wavelet filter/mother wavelet to use. The sparseness is a measure of how fast the DWT coefficients decay, and we are interested in obtaining a representation where the energy of the signal is concentrated in a few of the DWT coefficients. It is well-known that the decay of the DWT coefficients is strongly related to the number of vanishing moments of the mother wavelet, and to the smoothness of the signal. In this paper we present the result of applying two classical families of wavelets to a series of musical signals. The purpose is to determine a general relation between the number of vanishing moments of the wavelet and the sparseness of the DWT coefficients, when applied to music signals.

1. Introduction

The results presented are obtained as part of an ongoing research on automatic music classification. The idea is that by finding a sparse representation of music, good features that "capture the nature" of the music can be found. Achieving sparseness is not the main goal of the project, but a means to extract features that can be used for distinguishing different classes of music. One of the methods we investigate is representation of music signals in redundant dictionaries (see Section 1.1) containing a wavelet packet, and we want to base the choice of wavelet on numerical computations, within a solid theoretical framework.

1.1. Background

Many different methods for feature extraction from music or other sound signals exists, eg. [5], [7], and [9]. 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 categorizing into 2-3 classes, but the tests are performed on samples of very different size and content, and cannot be compared directly.

A music signal contains events of both short and long duration. At note onsets the amplitude of a number of frequencies grows rapidly (short duration event), and then decreases slowly (long duration). Singing also consists of both short and long duration events. Therefore the representation of a music signal in an orthonormal basis where the elements have similar structure (which is the case for Fourier and wavelet transforms) is not necessarily the most efficient in term of sparseness.

An alternative to orthogonal bases is decomposition in a more general dictionary, which is basically a collection of vectors (or waveforms) of the same length as the signal. The dictionary can consists of one or more bases, for instance the Fourier basis and the wavelet bases originating in a wavelet packet decomposition [12]. The elements in a dictionary are usually denoted atoms.

Our ultimate goal is to make efficient representations of music signals in dictionaries of various kinds. One of the key component in any dictionary, we believe, is wavelets. We therefore want to know which wavelets will be good for representing typical music signals. By a representation we mean a vector $\boldsymbol{x}$ which through the dictionary can reproduce the signal. More specifically, a sampled signal $\boldsymbol{b}$ of length n can be represented as

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

where $\boldsymbol{A}$ is an $n \times m$ full rank matrix containing the atoms in a dictionary as its columns, and $\boldsymbol{x}$ is the coefficients of the representation. For an orthonormal $\boldsymbol{A}$ the representation $\boldsymbol{x}$ is unique and $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$. When $m > n$ the representation is no longer unique, and we can choose among all possible representations the one

that is optimal for the application. The downside is that in general the complexity of finding the representation increases.

A optimal representation for feature extraction of music would be one for which the information is contained in as few coefficients as possible (when using a dictionary with atoms having well defined time, frequency and/or scale location). This corresponds to minimizing the ℓ^0 norm of the coefficients in (1). However minimizing the ℓ^0-norm is an NP hard problem, which is computationally expensive even for short signals. An alternative is to minimize the ℓ^1 norm, since it provides a representation where the main part of the energy is concentrated in a few coefficients, and at most n coefficients are non-vanishing (which is not the case for the ℓ^2 minimization given by the Moore Penrose inverse. The ℓ^1 minimization problem is also in general NP hard, but can be rewritten into a linear programming problem with smaller complexity [1].

1.2. Using Wavelets

In the present work we have investigated the approximation ability of various wavelets using just the DWT. This is admittedly in spite of the fact that we are interested in sparseness of ℓ^1 optimal solutions. The reason for this is that we want to initiate our investigation within a somewhat solid theoretical framework. Unfortunately, very few theoretical results exist (to the best of the authors' knowledge) on the relation between wavelet properties such as vanishing moments and regularity and the sparseness and decay of ℓ^1 optimal solutions. This – to some extent – also applies to best orthogonal bases obtained via the wavelet packet transform. Consequently, this initial investigation is carried out using the DWT for which the mentioned properties are well understood. We believe that this first, limited investigation will account for some of trends that we observe in a more general wavelet setting, and we will present related results in a later paper.

2. Methods

A series of normalized excerpts from music signals are wavelet transformed over the maximum number of scales using periodization at the ends of the signal. Fifteen different wavelet filters – having one through fifteen vanishing moments – from each of the two classical families Daubechies and symlets (least asymmetrical Daubechies wavelets, see [3]) are used for transformation. The filters, the filter length and the number of vanishing moments are presented in Table 1.

Table 1. The filter length and number of vanishing moments for the 30 different wavelets applied.

Wavelet family	Filter length	# vanishing moments
Daubechies	2, 4, 6, …, 30	1, 2, 3, …, 15
Symlet	2, 4, 6, …, 30	1, 2, 3, …, 15

The reason for investigating both of these two closely related families is to determine whether symmetry is significant for sparseness of music (as is the case for sparseness of images).

The DWT of a music excerpt $\boldsymbol{b}$ of length n can be represented as

$$\boldsymbol{A}^i \boldsymbol{x}^i = \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{A}^i$ is the $n \times n$ matrix representing the DWT using the i'th wavelet, and $\boldsymbol{x}^i$ is the coefficients of the transformation of $\boldsymbol{b}$. The sparseness measure is defined by

$$S(\boldsymbol{x}^i) = \frac{||\boldsymbol{x}^i||_1}{\frac{1}{30}\sum_{i=1}^{30} ||\boldsymbol{x}^i||_1}, \tag{3}$$

since it is not the level of the ℓ^1 norm that is interesting, but the ℓ^1 norm relative to the norms found for the same excerpt, but using other wavelets.

The ℓ^1 norm is used to measure the sparseness of the DWT coefficients. While this choice coincides with our interest in ℓ^1 norm as stated above, other sparseness measures might just as well have been used. For instance using Shannon's entropy produces more or less the same results.

Since we are looking for sparse representation we want a small ℓ^1 norm of the DWT coefficients. As argued in the next section this is more likely with longer wavelets, and we therefore anticipate to see that the ℓ^1 norm of the DWT coefficients decrease for increasing number of vanishing moments. At the same time an increased sparseness is linked to sufficiently high smoothness of the original signal, and since music signals in general are not particularly smooth we expect the norm to cease decreasing when the number of vanishing moments becomes too high. This will happen gradually as the number of vanishing moments increases because more and more of the signal energy is 'located in insufficiently smooth energy'.

This effect can also be used to estimate the smoothness of a musical signal. Due to the close link between smoothness of the wavelet and of the signal, and rate of decay of transform coefficients, we can conclude that as

the norm ceases to decrease the smoothness of the musical signal is found as the regularity of the corresponding wavelets. For the Daubechies and Symlets the regularity is approximately 0.5, 1, and 1.5 for 2, 3, and 4 vanishing moments, respectively [3].

2.1. A Note on Vanishing Moments

When searching for the best wavelet to use for representing music signals there are two opposing interests in respect to the length of the wavelet. That is to say, the number of filter taps in a discrete implementation. From a computational and numerical point-of-view we would prefer a short wavelet. Although a fast wavelet transform implementation is indeed quite fast, the methods for finding optimal representations in various settings are often iterative and thus need to apply the transform many times. The issue of numerical stability arises in fixed point and hardware implementations, which are the natural platform for musical analysis in low-cost consumer products.

From an approximation and sparseness point-of-view we want longer wavelets. Or more accurately, we want more vanishing moments and higher regularity, which is possible (but not implied) by using more filter taps. The desire for vanishing moments originates in the Strang-Fix condition which states (as a special case) that the approximation order of a wavelet transform increases with the number of vanishing moments of the wavelet up to the smoothness index (Hölder regularity) of the approximated signal [8, 10]. That is, the sparseness of the wavelet transformed signal is in general higher for longer wavelets.

While the number of vanishing moments is important for the sparseness (the decay of the sorted coefficients), the regularity of the wavelet is related to the decay of the unsorted transform coefficients. In fact, it is shown in [11] that wavelet expansions of sufficiently smooth functions converge at rates equal in measure to the differentiability of the wavelet. And conversely, we cannot expect better convergence when using wavelets with more smoothness than the approximated function.

The relation between vanishing moments and Hölder regularity (or Sobolev regularity, for that matter) is not simple. Although $\psi \in C^r$ does imply that ψ has $\lfloor r \rfloor$ vanishing moments (see e.g. [4, 2]), the converse is not true (and often far from). For instance, wavelets of the Daubechies family has $N/2$ vanishing moments, but belongs to $C^{\mu N}$, $\mu \approx 0.2$ for large N [3]. In fact, some constructions even sacrifice vanishing moments in order to gain freedom to increase the regularity, see for instance [6].

In this paper we have decided to use only vanishing moments as the 'sparseness potential' of each wavelet. Although this is only part of the truth, we believe it is sufficient for our purpose.

3. Results

The discrete wavelet transform is applied on signals sampled from 12 different pieces of music. From each song 30 consecutive excerpts are used. The excerpts have length 2^{13}, which corresponds to approximately 186 ms (the sample rate is 44.1 kHz). The sparseness measure $S(\boldsymbol{x}^i)$ for 30 signals from each of three songs are presented in Figure 1. The sparseness measure is plotted versus the number of vanishing moments. The measurements are represented as a quartile plot, each box has lines at the lower quartile, the median and the upper quartile. The whiskers indicate the extend of the data and outliers, which are more than 1.5 $\times$ interquartile range away from lower or upper quartile values, are marked with a *.

The results for the remaining 9 songs (not shown) are similar to the results for the "Jarre" and "Dion" songs, only one differs which is the "Accept" song, but the difference only lie in the level of the norms. For one vanishing moment the norms are spread out, but for higher number of vanishing moments, the variance is very small, indicating that the regularity of the signal relative to the regularity of the wavelet applied is the factor, that determines the sparseness.

3.1. Vanishing Moments and Regularity

Figure 1 shows that the sparseness measure $S(\boldsymbol{x}^i)$ do decrease as the number of vanishing moments of the mother wavelet increases. For the first two plots $S(\boldsymbol{x}^i)$ is about 20% higher for wavelets with one vanishing, than for wavelets with more than three or four vanishing moments, and for the last plot about 70% higher. So there is much to gain going from one to four or more vanishing moments. Three or four vanishing moments seem to be the limit where the regularity of the wavelet exceeds the regularity of the signal, and there is not much to gain in sparseness by increasing the number of vanishing moments. This indicates that the main part of the signal energy has regularity in the range 1 to 2.

3.2. Symmetry

The sparseness do not depend on which of the wavelet families are applied, the sparseness can be directly compared, since the ℓ^1 norm of the transformed signals are normed using the mean value of the ℓ^1 norms found for both the Daubechies and the symlet wavelets.

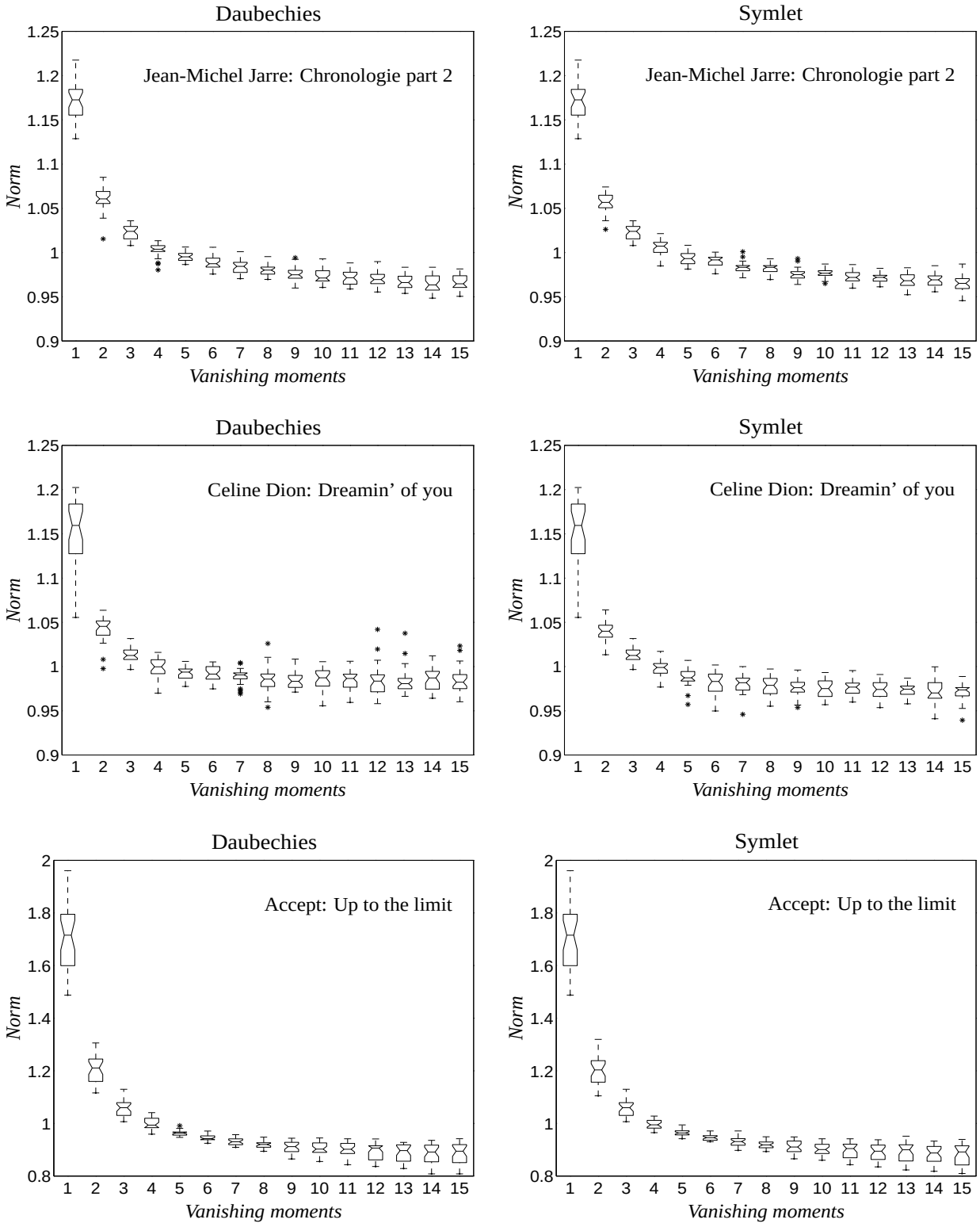

Figure 1. Sparseness measure applied to DWT coefficients. Each box with whiskers represents 30 consecutive excerpts of length 8192 from a music signal. The box itself extends from the lower quartile value through the median (middle notch) to the upper quartile value. The whiskers show the range of the remaining sparseness measures, except for outliers which are more than 1.5 $\times$ inter-quartile range away from lower or upper quartile values. Outliers are marked with *.

So the sparseness do not (for these two families of wavelets) depend on whether the wavelet is (close to) symmetric or not.

4. Discussion

The choice of wavelet to use for a given application is a challenge that arises all the time. In this work we have studied the importance of the number of vanishing moments for sparse wavelet representations of musical signals. The computations made on a series of music examples demonstrated that only a few vanishing moments are necessary for achieving a close-to-optimal sparseness. Although the test was carried out using only two types of wavelets a number of theoretical results indicates that other wavelet families would respond similarly (this can also be verified numerically), and thus that the number of vanishing moments and, to a lesser degree, the regularity, is important when choosing a wavelet for representing music signals.

The computations made also showed that the (Hölder) regularity of music signals is in the region of 1 to 2, and that only very little energy in the signals have higher regularity.

5. Acknowledgment

This work is in part supported by the Danish Technical Science Foundation, program no. 56-00-0143.

References

[1] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Comput.*, 20(1):33–61, 1998.

[2] I. Daubechies. *Ten Lectures on Wavelets*, volume 60 of *CBSM-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, Pa., 1992.

[3] I. Daubechies. Orthonormal bases of compactly supported wavelets. II. Variation on a theme. *SIAM J. Math. Anal.*, 24(2):499–519, march 1993.

[4] E. Hernández and G. Weiss. *A first course on wavelets*. CRC Press, Boca Raton, FL, 1996. With a foreword by Yves Meyer.

[5] S. Z. Li. Content-based audio classification and retrieval using the nearest feature line method. *IEEE Transactions on Speech and Audio Processing*, 8(5):619–625, September 2000.

[6] H. Ojanen. Orthonormal compactly supported wavelets with optimal sobolev regularity. *Applied and Computational Harmonic Analysis*, 10:93 – 98, 2001.

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

[8] G. Strang. *Numerical Analysis: A. R. Mitchell Birthday Volume*, chapter Creating and comparing wavelets. G. A. Watson and D. F. Griffiths, eds., World Scientific, 1996.

[9] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.

[10] M. Unser. Vanishing moments and the approximation power of wavelet expansions. In *Proceedings for International Conference on Image Processing*, volume 1, pages 629 – 632. IEEE, 1996.

[11] G. G. Walter. Approximation of the delta function by wavelets. *J. Approx. Theory*, 71(3):329 – 343, 1992.

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