Wavelet Bases for Compact Music Representation

Juan M. Vuletich and Ana M. C. Ruedin
Departamento de Computación,
Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires.
Email: jmvuletich@sinectis.com.ar, anita@dc.uba.ar

***Abstract*— We present a new basis for compact representation of music. The basis vectors are discretizations of a real Morlet wavelet, and they are induced by a tiling of the time-frequency plane well adapted to digital music. Through a careful investigation of their properties they are subsequently slightly modified in order to give an orthonormal system, without losing their time-frequency localization. This wavelet transform lowers the entropy of the musical signal, and enables compression if followed by an entropy based coder, such as arithmetic coding. Our new basis will also enhance the performance of all kinds of digital audio applications.**

I. Introduction

Music is made of sound waves. Having a basis for music that gives information both on the time and and the frequency of the notes that are played, will give an optimal representation of the signal. When the same note is held for several small time-intervals, there will be repetitions of the coefficients, and this will lower the entropy. Therefore, finding good bases for joint time-frequency representation of music is a key problem.

The construction of bases that distinguish musical notes, requires that their spectra be mainly supported on frequency intervals that have a constant relative bandwidth, and this constant is irrational. Also the support (in time) of these bases must vary with the frequency, to have good resolution in time.

Different transforms have been employed for dealing with digitized music. Continuous transforms such as the Fourier Transform, the Windowed Fourier Transform or Gabor Transform [1], and the Continuous Wavelet Transform (CWT), which are appropriate for signal analysis, have to be discretized in order to be applied to a discrete signal. They generally provide redundant representations of the signal in terms of frames, therefore they do not perform well for coding the signal. See [2] for an overview.

The Discrete Fourier Transform (DFT) and the Discrete Cosine Transform (DCT) have no time localization. The Modified Discrete Cosine Transform (MDCT) used in MPEG audio compression [3] has some time information; the vector bases are oscillations modulated by an envelope, as in the Windowed Fourier Transform. Yet the bandwidths and the time intervals are of fixed length. On the other hand, the Discrete Dyadic Wavelet Transform (DWT) [5] has good time localization, but its frequency localization is poor; it is not good enough to discriminate semitones, only octaves. Wavelet packets [6] improve the frequency resolution, but cannot discriminate semitones.

Because of these drawbacks, several attempts have been made to make more general partitions of the time-frequency plane. In [7], orthonormal bases are constructed for general tilings. Reduction of a dictionary of multiple Gabor frames is proposed in [8] and [9].

We specifically focus on tilings suitable for music representation, and present a new orthonormal basis, that has good time and frequency localization, for compact representation of music. In sections II to IV we explain the construction of the basis, which was outlined in the first author's earlier work [10] and grade thesis [11]. In section II we make a partition or tiling of the time-frequency plane that reflects the construction of the musical scale. In section III we define the elements (or atoms) of the basis (which are discretizations of a real Morlet wavelet), one for each tile, having most of their energy concentrated on that tile. Subsequent modifications of the tiling, and orthogonalization will modify our basis (section IV). In section V we present some preliminary tests, and our concluding remarks in VI.

II. An ideal tiling of the time-frequency plane

The pitch or frequency of a note is measured in cycles per second, or Hertz (Hz); for example the note central A (La central) has 440 Hz. Our perception of pitch is logarithmic: to our ear, the three notes 220 Hz, 440 Hz and 880 Hz sound equally spaced apart, yet their frequencies are related by a multiplicative factor. When the frequencies of two notes differ by a factor of 2, as in this case, we have an octave. In the twelve-tone equal-tempered scale there are 12 notes or semitones per octave, and the frequency ratio of 2 adjacent notes is constant: we call this constant a_0. Take A as the first note of an octave: then the second note, a semitone above A, will have frequency $440\ a_0$ Hz, the third $440\ a_0^2$ Hz, and so on. The thirteenth note, an octave above A, will have frequency $440\ a_0^{12} = 440 \times 2$ Hz. This means that $a_0 = \sqrt[12]{2}$, an irrational number.

Suppose the original signal has been sampled at a rate of F_s samples per second. The maximum frequency is $\frac{F_s}{2}$ Hz.

First, the time-frequency plane is divided into horizontal bands, each one covering a different frequency interval. Then each band is sliced in time, independently of the others, into rectangles of the same length.

Figure (1) shows an ideal ideal tiling of the time-frequency plane for an octave. There are twelve bands, one for each semitone.

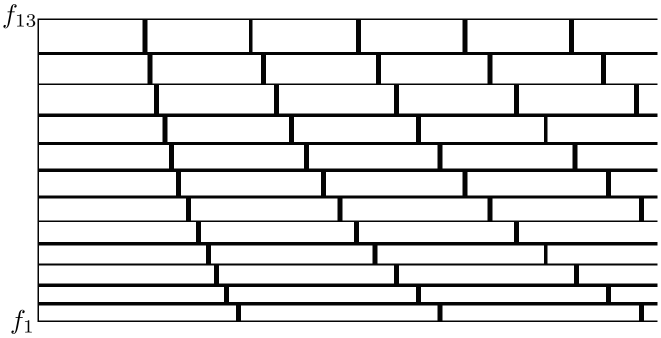

Fig. 1. Tiling of the time-frequency plane for an octave.

To make our construction simple, we concentrate on an octave. We divide the frequency axis into intervals or bands $[f_j, f_{j+1}]$, $j = 1 \ldots J$, where $J = 12$. We want the frequencies corresponding to the 12 semitones to be the midpoints of the bands.

Let f_1 be the lower limit of the band for the first note, we shall later deduce its value. We build the sequence $\{f_j\}_{j=1\ldots J+1}$ that limits the frequency bands as

$$f_{j+1} = a_0 \ f_j. \tag{1}$$

Let $f_j^{\,c}$ be the central frequency of band $[f_j, f_{j+1}]$:

$$f_j^{\,c} = \frac{f_j + f_{j+1}}{2},$$

and let $\Delta f_j = f_{j+1} - f_j$ be the bandwidth.

It can be easily verified that the geometric progression property (1) of $\{f_j\}$ is inherited by its middle points and by its differences, giving

$$f_{j+1}^{\,c} = a_0 \ f_j^{\,c} \quad (i); \qquad \Delta f_{j+1} = a_0 \ \Delta f_j \quad (ii). \tag{2}$$

We want the lowest central frequency $f_1^{\,c}$ to be exactly the frequency of the first note of the octave. By construction (equation (2)(i)) , the remaining semitones will have frequencies $f_j^{\,c}$, $j = 2 \ldots 12$.

Since

$$f_j^{\,c} = \frac{(1+a_0) f_j}{2} \quad (i), \quad \text{then} \quad f_j = \frac{2 \ f_j^{\,c}}{1+a_0} \quad (ii), \tag{3}$$

and substituting $j = 1$ in (3)(ii), we may calculate f_1.

Although the bandwidth $\Delta f_j = (a_0 - 1) f_j$ varies with j, the relative bandwidth is constant

$$\frac{\Delta f_j}{f_j} = a_0 - 1 \quad (i), \quad \text{and} \quad \frac{f_j^{\,c}}{\Delta f_j} = \frac{(a_0+1)}{2(a_0-1)} \quad (ii). \tag{4}$$

There is no fixed division for the time axis as for the frequency axis. Tiles belonging to the same frequency band have equal lengths. Let $[\ t_{j,k}, t_{j,k+1}]$ be an interval in time for the frequency band j, i.e. for frequencies that lie in $[f_j, f_{j+1}]$. Then the time-interval or width of the tile is $\Delta t_j = t_{j,k+1} - t_{j,k}$.

Heisenberg's uncertainty principle states that it is not possible to have both time and frequency (perfect) localization [12], [4]. The area of each tile has an inferior bound; the tiles we will construct will all have the same area. Accordingly, we set

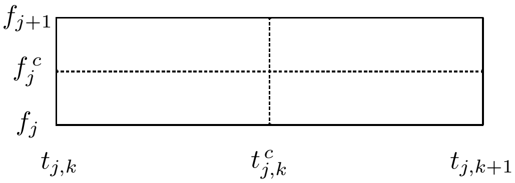

Fig. 2. A tile (j,k)

$$\Delta f_j \quad \Delta t_j = \frac{1}{2}. \tag{5}$$

For low-pitched sounds, the frequency bandwidth is smaller than for high-pitched ones. As expected, a greater time interval is needed to identify the frequency of a signal that has fewer cycles per second.

Once Δt_j is obtained from (5), we can calculate $t_{j,k+1} = t_{j,k} + \Delta t_j$. We set $t_{j,0} = 0$ for all bands. Then $t_{j,k} = k \ \Delta t_j$.

Each tile (j,k) –see figure (2)– is centered at frequency $f_j^{\,c}$ and at time

$$t_{j,k}^{\,c} = \frac{t_{j,k} + t_{j,k+1}}{2} = \left(k + \frac{1}{2}\right) \ \Delta t_j.$$

III. Construction of the basis

Once the tiling of the time-frequency plane is achieved, the next problem is to find a basis that is well localized over such tiles. We have chosen the Morlet wavelet, a modulated Gaussian, that has often been used for music analysis with CWT [13] because it reaches the theoretical limit to time and frequency localization specified by Heisenberg's uncertainty principle.

The Morlet wavelet is a complex function; since our signal is real, we have chosen its real part:

$$\Psi(t) = b \ \sqrt{\pi} \ e^{-(b \ \pi \ t)^2} \cos{(2 \ \pi f \ t)}, \tag{6}$$

where parameter b controls the width of the Gaussian.

Appropriate dilations and displacements of the wavelet will allow us to place it over any tile of the partition, to obtain all the elements of the basis. However, it is impossible to confine a wavelet strictly to a tile, because a function cannot be compactly supported both in time and frequency domain. We aim at having most of the energy of the wavelet concentrated on a tile, and have good decay on the neighbouring tiles.

It is however possible to choose a better localization in one domain, accepting a worse localization in the other. Here frequency localization is priviledged over time localization, to allow good discrimination of the semitones. This decision is made to match the characteristics of our hearing system.

It can be proven that the Fourier Transform of the wavelet is the mean of two Gaussian functions, one centered at $u = f$ (which is of interest) and the other centered at $u = -f$ (of no interest):

$$\widehat{\Psi}(u) = \frac{1}{2}\ \left[e^{-(\frac{u-f}{b})^2} + e^{-(\frac{u+f}{b})^2}\right], \tag{7}$$

where the The Fourier Transform of a function $x(t)$ is defined as

$$\widehat{x}(u) = \int_{-\infty}^{\infty} e^{-\,i\,2\,\pi\,u\,t}\ x(t)\,dt. \tag{8}$$

It is clear from formulae (6) and (7) that parameter b controls the balance between time and frequency localization. To priviledge frequency localization over temporal localization, we chose $b = 0.31\,\Delta f_j$.

To construct our basis, we place a wavelet on the center of each tile. For each tile (j,k) we have the corresponding wavelet $\Psi_{j,\,k}(t)$, having central frequency $f = f_j^{\,c}$, and centered at time $t = t^{\,c}_{j,\,k}$:

$$\Psi_{j,k}(t) = b\,\sqrt{\pi}\,e^{-(b\,\pi\,(t-t^{\,c}_{j,\,k}))^2}\cos\left(2\,\pi f_j^{\,c}\,(t-t^{\,c}_{j,k})\right).$$

Figure (3) shows two consecutive displacements of the wavelet, and their product.

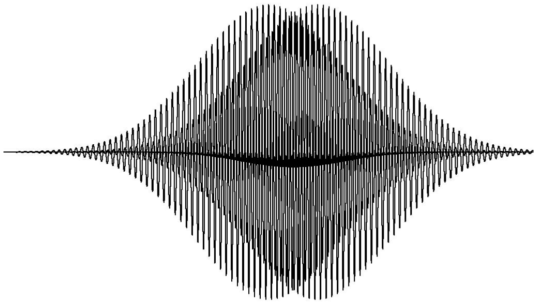

Fig. 3. Two neighbouring wavelets $\Psi_{j,\,k}$,$\Psi_{j,\,k+1}$ and their product.

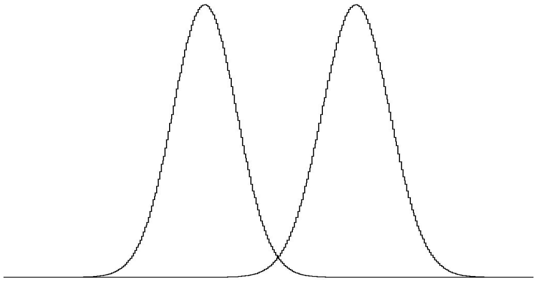

Fig. 4. Spectra of 2 Morlet wavelets $\Psi_{j,\,k}$,$\Psi_{j+1,\,k}$ on consecutive bands.

In figure (4) are shown the spectra of two wavelets belonging to neighbouring bands, i.e. having central frequencies $f_j^{\,c}$ and $f^{\,c}_{j+1}$ corresponding to two consecutive semitones: observe the very small overlap.

The parameter t of the wavelets $\Psi_{j,\,k}$ is a continuous variable: the next step is to discretize the wavelets, by evaluating them on values of t $\frac{1}{F_s}$ second appart, like the samples of the original signal, thereby obtaining vectors $V^{\,(j,k)} \in \Re^N$ of coordinates

$$V_n^{(j,k)} = \Psi_{j,\,k}\left(\frac{n}{F_s}\right), \qquad \text{for} \qquad n = 1\ldots N,$$

supposing the total time under consideration to be $\frac{N}{F_s}$ seconds.

IV. Orthogonalization of the basis

At this point our basis consists of vectors $V^{\,(j,k)}$, ($j = 0,\ldots J+1$) of length N. In all there should be N tiles and N basis vectors. We normalize each vector, and place them as the columns of a matrix A. (We can number the tiles of the plane according to the values of $t_{j,k}$, and follow that order to place the basis vectors in A.) The large dimension of the matrix makes it desirable to have an orthonormal basis: this will allow easier and faster computation of the transform and the reconstruction of signals. To run an orthogonalization process like Gram-Schmidt [14] on the matrix would make havoc in the time-frequency localization of the basis vectors. Therefore, it is necessary to proceed carefully.

A. Correlation between elements of different bands

Because of Parseval's equation, we know that the correlation between 2 basis vectors is equal to the correlation of their FFT's. This indicates that basis vectors belonging to consecutive bands may not be orthogonal – in fact, the correlation is very small, because of the small overlap of their spectra–, but it also indicates that basis vectors belonging to bands farther appart are indeed orthogonal.

B. Correlation between elements of the same band

A way to reduce the correlation of wavelets in the same band, is to make the small oscillations of the wavelet to be at a phase of $90°$ in neighboring displacements. The frequency of these small oscillations is the central frequency $f_j^{\,c}$ of the tile. Therefore Δt_j, the width of the tile measured in seconds, must be an integer number of cycles of frequency $f_j^{\,c}$ plus $\frac{1}{4}$ or $\frac{3}{4}$ of a cycle, i.e. the number of cycles should be $\frac{L}{2}+\frac{1}{4}$, with integer L. Since there are $f_j^{\,c}$ cycles per second, there are $\frac{2L+1}{4}$ cycles in $\frac{2L+1}{4f_j^{\,c}}$ seconds.

This means that $\Delta t_j = \frac{2\,L+1}{4\,f_j^{\,c}}$. Recall that $\Delta t_j = \frac{1}{2\,\Delta f_j}$, and substitute Δf_j from equation (4). We get

$$\frac{2\,L+1}{4\,f_j^c} = \Delta t_j = \frac{1}{2\,\Delta f_j} = \frac{(a_0+1)}{4\,(a_0-1)\,f_j^{\,c}},$$

from which we obtain $a_0 = 1 + \frac{1}{L}$.

At this point, the orientation of this research needed to be changed. From a true musical scale tiling of the plane where the ratio of two consecutive limitting frequencies was equal to an irrational number a_0 –see equation (1)–, the focus was shifted to a tiling of the plane where the ratio of two consecutive limitting frequencies is equal to the best rational approximation $1+\frac{1}{L}$ of a_0.

The fractions of the form $1 + \frac{1}{L}$ closer to $a_0 = \sqrt[12]{2} \approx 1.05946$ are $1 + \frac{1}{17} \approx 1.05882$ and

$1 + \frac{1}{16} \approx 1.0625$. We can therefore approximate the 12 bands of an octave by 10 slightly narrower bands and 2 slightly wider bands : instead of having ${a_0}^{12} = 2$, we have

$$\left(1 + \frac{1}{17}\right)^{10} \left(1 + \frac{1}{16}\right)^{2} \approx 1.9993725.$$

The error is negligible.

With this new tiling, we have orthogonality between all displacements at odd distances of the wavelet in the same band j, for $1 \leq j \leq J$. In fact, it can be proven that

$$\int_{-\infty}^{\infty} \Psi_{j,k}(t)\Psi_{j,k+2m+1}(t)dt = 0;$$

it follows that $V^{(j,k)}$ and $V^{(j,k+2m+1)}$ are orthogonal, if the discretization is sufficiently fine.

m	±2	±4	±6	±8	±10
	-0.62	0.15	-0.014	0.0005	-0.0000071

TABLE I

SCALAR PRODUCT OF $V^{(j,k)}$ AND $V^{(j,k+2m)}$

C. *Final orthogonalization of the basis vectors*

After having reduced the correlation between basis vectors, it remains to orthogonalize the whole basis. This is done in 2 steps. First, in an iterative process, we select the higher correlation between remaining basis vectors at even distances in the same band, say $2n$, and subtract the projection of one of the vectors (multiplied by an carefully choosen factor between 0 and 1) from the other, at distances $2n$ and $-2n$, in order to maintain symmetry of the basis. Second, we orthonormalize the whole basis.

V. TESTS

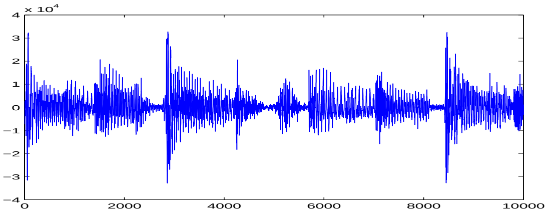

Fig. 5. Original signal.

Preliminary results are encouraging. We tested our algorithm on a short recording extracted from the theme “No tan distintos” (Not so different) of the Argentine Rock Band “Sumo”, a melody of 6 notes played by an electric bass guitar (see figure 5). The original signal has 10000 samples. It was coded at a rate of 5512.5 samples per second, with 16 bits per sample, the samples being in the range $[-32768, 32767]$.

We constructed a basis covering 3 octaves, that is 36 bands. We added 2 extra bands, one for lower frequencies and the other for higher frequencies; in all there are 38 frequency bands. In figure (6) are some elements of the basis belonging to the same time interval and different frequencies, and their spectra.

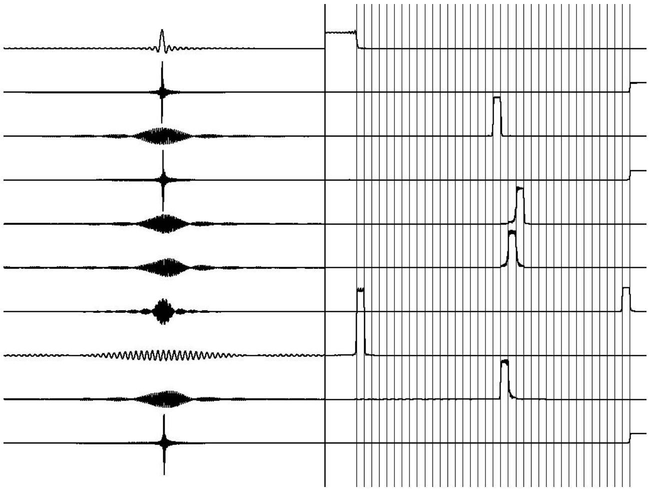

Fig. 6. Some elements of the basis and their spectra.

We have calculated the coefficients of this signal in terms of our basis, by simply multiplying our signal by the transpose of matrix A. We show the coefficients in a grey-scale map – figure 7 above–, where the darker shades of gray mean higher absolute values. Observe that most of the coefficients are light-shaded, i.e. small. The darker-shaded groups of coefficients, around which we have drawn ellipses by hand, are in correspondence with the fundamental notes played (see the music score in the figure, below) and their overtones.

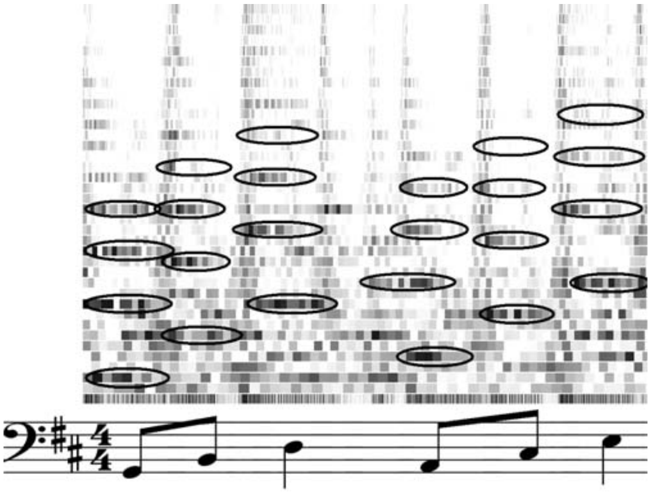

Fig. 7. Map of coefficients (above) and music score (below).

The coefficients are floating numbers in the interval $[-106972, 112985]$, a wider range than the signal samples. To reduce them to 16 bits, they were quantized (divided by 3.36 and rounded), which for each coefficient represents an error –relative to the largest coefficient– at most 0.000016. In table II we may observe how the proposed tranform lowers the entropy of the original signal. This will result in higher compression if it is followed by arithmetic coding [15]. To recover the signal, we multiply the coefficients by A.

Original signal samples	Entropy 12.72	n^o different symbols 7439
Quantized coefficients	Entropy 10.35	n^o different symbols 2964

TABLE II

ENTROPY OF ORIGINAL SIGNAL AND OF COEFFICIENTS IN OUR BASIS

VI. CONCLUSION

We have proposed a new basis for compact music representation. The basis vectors are a modification of a discretized real Morlet wavelet, they are induced by a tiling of the time-frequency plane well adapted to digital music, and they are orthonormal.

We have tested our transform, and preliminary results show that most of the coefficients in our base are small, indicating that their histogram is peaked at zero. There are less different coefficients, and the entropy of the transformed signal is lower than the entropy of the original signal. All this indicates their suitability for music compression.

REFERENCES

[1] J. Alm and J. Walker, “Time-frequency analysis of musical instruments,” *SIAM Review*, vol. 44, no. 3, pp. 456–476, 2002.

[2] B. Torrésani, “An overview of wavelet analysis and time-frequency analysis,” *Proceedings of the International Workshop in Self-Similar Systems (Dubna, Russia)*, p. 22, 1998.

[3] D. Pan, “A tutorial on mpeg/audio compression,” *Multimedia, IEEE*, vol. 2, pp. 60–74, 1995.

[4] I. Daubechies, *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992.

[5] M. V. Wickerhauser, *Adapted Wavelet Analysis : FromTheory to Software*. A K Peters, 1994.

[6] R. Bernardini and J. Kovacevic, “Arbitrary tilings of the time-frequency plane using local bases,” *IEEE Transactions on Signal Processing*, vol. 47, no. 8, pp. 2293–2304, 1999.

[7] M. Dorfler, “Gabor analysis for a class of signals called music,” Ph.D. dissertation, Universitat Wien, July 2002.

[8] M. Dorfler and H. Feichtinger, “Quilted gabor families I: Reduced multi-gabor frames,” *submitted to Elsevier Science*, vol. preprint, 2005.

[9] J. M. Vuletich, “Orthonormal bases and tilings of the time-frequency plane for music processing,” *Proc. of SPIE, Wavelets: Applications in Signal and Image Processing X; Michael Unser, Akram Aldroubi, Andrew F. Laine, Editors*, vol. 5207, pp. 784–793, 2003.

[10] ——, “Nuevas bases para el procesamiento de música en el dominio tiempo-frecuencia,” *Tesis de Licenciatura, Univ. de Buenos Aires, Dirección: A. Ruedin*, Abril 2005.

[11] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[12] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley Cambridge Press, 1996.

[13] G. Olmo, F. Dovis, C. Calosso, and P. Passaro, “Instrument independent analysis of music by means of the continuous wavelet transform,” *Proceedings SPIE Wavelet Appl. Signal Image Proc. VII*, vol. 3813, pp. 716–723, 1999.

[14] G. Golub and C. Van Loan, *Matrix Computations*. North Oxford Academic Publishers, 1986.

[15] T. Cover and J. Thomas, *Elements of information theory*. Wiley-Interscience, 1991.