INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

TU Delft
Delft University of Technology

MULTIDIMENSIONAL COMPANDING

A Multidimensional Companding Scheme for Source Coding with a Perceptually Relevant Distortion Measure

João Bernardo Farinha Pereira Crespo

Dissertação para obtenção do Grau de Mestre em[1]

Engenharia Electrotécnica e de Computadores

Júri

Presidente: Prof. Dr. José Manuel Bioucas Dias

Orientador[2]: Prof. Dr. Pedro Manuel Quintas Aguiar

Vogal: Prof. Dr. Mário Alexandre Teles Figueiredo

Novembro 2009

[1] Tese feita em parceria com a Technische Universiteit Delft, Holanda

[2] Co-orientador: Dr. ir. Richard Heusdens

Special Thanks

Hereby, I would like to mention and thank a number of persons and institutions, who directly or indirectly contributed to the development of my thesis. They contributed to enriching the ideas present in the thesis and to the ways of expressing them.

In the first place, I would like to acknowledge and promote the whole infrastructure that makes the *Lifelong Learning Programme – Erasmus* exchange program possible. Indeed this program provided me a marvelous experience of living in another country (the Netherlands in particular), making it possible for me to enjoy the cultural differences between northern west-european and southern west-european cultures.

Concerning this thesis more specifically, I would like to thank my supervisors Richard Heusdens and Pedro Aguiar for supporting me during the whole process of it, starting from the theoretical work and ending in the writing part. Richard was my supervisor in Delft, and he was always very friendly and flexible in guiding me when I did not know what to do next and answering my questions. Although Pedro was at a somewhat large distance in Lisbon, he showed interest in the work and helped me to reformulate the abstract and introduction in a clear, concise and very well structured way. I thank you both for the help provided during this last period of my master program.

Furthermore, Jorge Martinez and Cees Taal, members of the Audio Lab in Delft, contributed an indispensable help for clarifying topics related to S. van de Par's perceptual distortion measure and to methods for solving complex differential equations, respectively. I would like to thank you both for those contributions, and emphasize that they played an important role in my thesis. Without your help, the current outcome of the thesis would not have been possible, and the thesis would have taken a completely different (much less nicer) course.

Now on the non-technical realm, I would like to praise the nice atmosphere that I experienced at TU Delft, namely in the ICT group, in the Audio Lab and with the group of master students. It was nice either to go out on a group trip, on a dinner, on a drink at the /Pub, or simply to talk for a while after (or during!) a day of hard and exhausting work (cof, cof![1]), to relax a bit in an informal environment.

Last but not least, I would also like to thank the not less important supporting role of my family, who always wished me good luck for the accomplishment of my thesis and welcomed me with open arms on my trips to Portugal.

[1] I think I caught the *Mexicaanse griep*!

Abstract

This thesis deals with the topic of multidimensional companding audio source coding. In this type of source coding, the vector source is passed by a pre-processing function, which we call the *compressor*, by a vector quantizer and finally, as a post-processing step, by the inverse of the pre-processing function, which we call the *expander*. Optimal multidimensional companding has the characteristic that locally quadratic distortion measures get mapped into the Mean Square Error (MSE) distortion measure in the compressed domain. Multidimensional companding enables thus the efficient usage of quantization schemes designed for the MSE with locally quadratic distortion measures, if we use the optimal compressor and expander.

Recently, S. van de Par et al. [59] developed a locally quadratic perceptual distortion measure for sinusoidal audio coding. In that work, the distortion is computed using a weighted MSE in the frequency domain, where the weights are given by the inverse of the masking threshold, a measure of the time-frequency dependent sensitivity of the human ear. This distortion measure has been successfully employed in several audio coding schemes, such as [58], [34], and [19].

In this thesis, we combine the technique of multidimensional companding with the mentioned perceptual distortion measure. The main contribution is the development of a multidimensional compander (compressor and expander), which is asymptotically optimal in the sense that it has a vanishing rate-loss with increasing vector dimension. The compressor operates in the frequency domain: in its simplest form, it point-wise multiplies the Discrete Fourier Transform (DFT) of the windowed input signal by the square-root of the inverse of the masking threshold, and then goes back into the time domain with the inverse DFT. The expander is based on numerical methods: we do one iteration in a fixed-point equation, and then fine-tune the result using Broyden's method.

Additionally, we show simulations which corroborate the approximations and results of the theoretical derivations.

Keywords: Multidimensional companding, locally quadratic distortion measure, perceptual distortion measure, sinusoidal audio coding

Resumo

Esta tese lida com o tópico da codificação de fontes áudio usando compansão multidimensional. Neste tipo de codificação de fonte, a fonte (um vector) é passada por uma função de pré-processamento, à qual chamamos *compressor*, por um quantizador vectorial e finalmente, como passo de pós-processamento, pela função inversa, denominada por *expansor*. Compansão multidimensional tem a característica de transformar medidas de distorção localmente quadráticas no erro quadrático médio (MSE), no domínio do sinal comprimido. Consequentemente, a compansão multidimensional possibilita o uso eficiente de esquemas de quantização desenhados para o MSE com medidas de distorção localmente quadráticas, se se usar o compressor e expansor óptimos.

Recentemente, S. van de Par et al. [59] desenvolveram uma medida de distorção perceptiva e localmente quadrática para codificação de áudio sinusoidal. Nesse trabalho, a distorção é calculada usando um MSE ponderado no domínio da frequência, onde os pesos são dados pelo inverso do limiar de mascaramento, uma medida da sensibilidade do ouvido humano, dependente do tempo e da frequência. Esta medida de distorção foi aplicada com sucesso em vários esquemas de codificação de áudio, como por exemplo em [58], [34], e [19].

Nesta tese combina-se a técnica da compansão multidimensional com a referida medida de distorção perceptiva. A contribuição principal é o desenvolvimento de um compansor multidimensional (compressor e expansor) assimptoticamente óptimo, no sentido em que o débito adicional produzido em relação ao débito do esquema óptimo desaparece ao incrementar a dimensão vectorial. O compressor opera no domínio da frequência: na sua forma mais simples, este multiplica ponto-a-ponto a transformada discreta de Fourier (DFT) do sinal de entrada, por sua vez multiplicado ponto-a-ponto por uma janela, pela raiz quadrada do inverso do limiar de mascaramento, e volta para o domínio do tempo com a DFT inversa. O expansor é baseado em métodos numéricos: é executada uma iteração numa equação do ponto fixo, e o resultado é ajustado usando o método de Broyden.

Adicionalmente, são executadas simulações que corroboram as aproximações e resultados da parte teórica.

Palavras Chave: Compansão multidimensional, medida de distorção localmente quadrática, medida de distorção perceptiva, codificação de áudio sinusoidal

Contents

List of Tables

List of Figures

List of Abbreviations

AAC	Advanced Audio Coding
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
HILN	Harmonic and Individual Lines plus Noise
iDFT	Inverse Discrete Fourier Transform
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
LVQ	Lattice Vector Quantization/Quantizer
MPEG	Moving Picture Experts Group
MSE	Mean Square Error
PAMP	Psycho-acoustical Matching Pursuit
PCM	Pulse Code Modulation
RAM	Random Access Memory
SiCAS	Sinusoidal Coding/Coder of Audio and Speech
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SSC	SinuSoidal Coding
WWW	World Wide Web

Chapter 1

Introduction

1.1 Initial Considerations

In the last decade, we have observed an explosive increase in the usage of audio coding schemes and coded audio content, which enable the reduction of the information throughput (bit-rate) of an audio signal on the order of 7 to 15 times with respect to the original Pulse Code Modulation (PCM) coded signal, with very reduced penalty in perceptual quality [56]. These schemes make countless applications possible, such as handheld audio decoders with reduced memory capacity, which nevertheless can carry hours of audio content, streaming through bandwidth constrained channels, such as the Internet, with low bandwidth usage and high experienced quality, delivery of audio content interactively through the World Wide Web (WWW), digital radio, digital television, recorded digital video, and many more. This decrease in bit-rate, yet with surprisingly high transparency, is achieved through the exploitation of the perceptual irrelevance and statistical redundancy present in the audio signal [46]. Indeed, certain time-frequency components of the signal are irrelevant to humans, since the time varying, signal dependent nature of the sensitivity of the human ear makes it possible for the distortion obtained from not transmitting those signal components to be masked by strong components of the signal in neighboring frequency bands or time instants. The statistical redundancy is inherent to the fact that not all signal samples carry the same amount of information, due to a non-uniform distribution of their probability of occurrence and to the existence of correlation between neighboring samples.

Contemporary audio coders can be subdivided in *transform coders* and *parametric coders.* Both types do a psycho-acoustical analysis on the input signal, having as output the so-called *masking threshold*, a time-frequency function that delivers the maximum amount of quantization noise, which is insertable in the source signal for each time-frequency interval, such that the distortion in the coded version is still inaudible. In parallel, either a linear transform is applied to a block of the signal (transform coders) or perceptually relevant parameters are extracted from it (parametric coders). Examples of transform coders are the ISO/IEC MPEG-1 Audio Layer 3 [1] and the ISO/IEC MPEG-4 Advanced Audio Coding (AAC) [2] coders, which are based on the MDCT transform [40]. As to the parametric coders, we have the Harmonic and Individual Lines plus Noise (HILN) [3], the SinuSoidal Coder (SSC) [4] and the Sinusoidal Coder of Audio and Speech (SiCAS) [19], which model the input signal as a sum of sinusoids and noise. Either way, the result is time-frequency dependent information of the input signal. This information is then quantized with accuracy compatible with the calculated masking threshold: the higher the masking threshold in a certain time-frequency bin is, the lower is the sensitivity of the human ear in that time-frequency region, and thus the higher the quantization step size (i.e., the coarser the quantization) is allowed to be, and vice-versa. In some coders, a residual signal is built from the difference of the original and the reconstructed signals and additional parameters extracted. Finally, the resulting symbols and side information are entropy coded, multiplexed, sent through the channel and the inverse procedure (demultiplexing, entropy decoding, inverse quantization and re-synthesis) is done. See [46] and [18,50] for an overview on transform and parametric audio coding, respectively.

In the quantization block of the encoder described above, it is desirable to quantize the information extracted from the signal in a rate-distortion optimal sense, i.e., in a way that minimizes the perceptual distortion experienced by the user subject to the constraint of a certain available bit-rate. That optimality is described by the rate-distortion function $R(D)$ [9], which expresses the minimum theoretically possible bit-rate R that we can achieve when coding at an expected distortion level not larger than D. This function does only depend on the source (the extracted information described in the previous paragraph) and on the *distortion measure* d that we use to quantify the (perceptual) discrepancy between the source x and the reproduction y (the signal at the decoder). At a rate $R(D)$, we have thus $\mathrm{E}\ d(x,y) \leq D$, where E denotes the expected value. Although very powerful, this function is only known completely for a reduced number of combinations of source and distortion measure, and in specific conditions with specific assumptions [65], being the most simple example the Gaussian source with the *mean square error* (MSE), $D = \mathrm{E}\ \|x - y\|^2$, as the distortion criterion [14]. Other examples include the Gaussian source with a weighted mean square error distortion measure [52], an arbitrary source with well-behaved source density using a generic difference distortion measure (distortion of the form $d(x,y) = \rho(y - x)$) at high resolution ($D \to 0$) [36], and an arbitrary source with a locally quadratic distortion measure (i.e., expressible as a quadratic form when $y \to x$), also assuming regularity conditions and good behavior of the source [37].

1.2 Motivation

Due to its mathematical tractability, the MSE is the elected choice for the distortion measure in many coding applications. In audio coding, however, the usage of more complex perceptual distortion measures, different from the MSE, exploiting the frequency selectivity and masking phenomena of the human auditory system, achieves a much higher performance, i.e., a higher perceptual quality for the same bit-rate. Due to the difficult mathematical tractability of these measures, the complete rate-distortion function is not known, and the direct design of rate-distortion optimal quantizers, for these measures, is mathematically intractable as well. A way to handle these measures indirectly is to multiply the input signal by certain perceptual weights related to the masking threshold before quantizing on the encoder and do the inverse (divide) on the decoder, so that the perceptual distortion measure gets mapped into an MSE in the normalized domain. The ease of use of the MSE distortion measure makes this solution attractive, having been employed in several quantization schemes, both in transform coding [2], [17], [65], and in sinusoidal coding [58], [34]. However, such an approach has the inconvenient that the weights have to be transmitted as side information through the channel so that the receiver can do the inverse normalization, thereby introducing an overhead in the transmission process.

The overhead may be intolerably high in certain contexts. Imagine the scenario of multiple description coding [22], represented in Figure 1.1. In this scenario, the transmission channel is modeled by a *packet erasure channel* with a certain number n of "sub-channels" ($n > 1$), where the so called *descriptions* (representations) of the input signal are transmitted. The information is coded at the source in n different ways; one stream for each description. In each transmission, for each description, this channel either lets the information (the "packet") pass through error-free or it suppresses ("erases") the packet completely. The packets which were erased are uncertain at the source. We have thus $m \leq n$ received packets, and the function of the receiver is to reconstruct the source with all the received information, in such a way that the higher the number of received packets is, the better the quality we achieve. This contrasts with the single description case in which $n = 1$ and $m \in \{0, 1\}$, where we either receive the information at full quality or we do not receive anything.

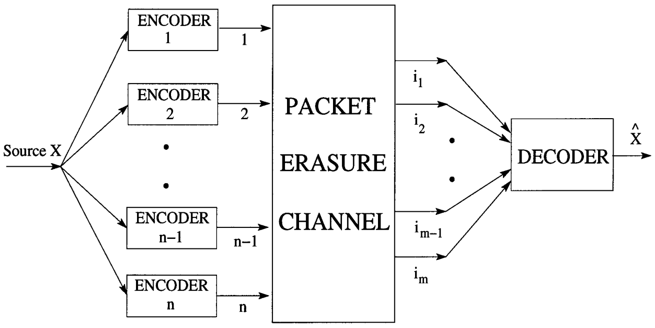


Figure 1.1: A Multiple Description Coding Scenario. Source: [49].

In a multiple description audio coding scenario, the side-information (the perceptual weights) relative to the masking threshold have to be transmitted in all packets, since we do not know at the transmitter which packets will arrive. In the worst case[1], $m = 1$, and we do not know which specific channel did not erase the packet. As that information takes a fixed amount of bits to code, not dependent on the number of descriptions (e.g., 4 kbps in [43]), for a fixed total rate, the more descriptions there are, the less usable rate there is for the real audio information (and the higher the percentage of usage is of the side-information). Therefore, from a certain point on, the performance of such a multiple description scheme degrades for increasing n. This problem was faced in [65], where the author developed a well-performing multiple description coding scheme designed for the MSE distortion measure, adapted for audio coding using the mentioned normalization by perceptual weights.

To solve this increase of overhead with n, we will explore the theory of *multidimensional companding* [38] in this thesis. In a multidimensional companding quantization scheme (see Figure 1.2), the source x, a vector (e.g., built from packing together consecutive signal samples), is first pre-processed by applying a non-linear vector function F, which we call the *compressor*, then a vector quantization source coding scheme is applied (quantization, entropy coding, transmission, entropy decoding and de-quantization), and finally the inverse function F^{-1}, the *expander*, is applied to the received signal as a post-processing step. The set of both functions, the compressor and the expander, is called the *compander*.

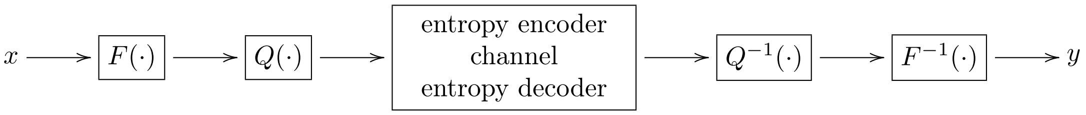


Figure 1.2: Source coding with multidimensional companding.

In [38], the authors show that general non-difference, locally quadratic distortion measures (which satisfy some technical conditions) are equivalent to the MSE measure at high resolution in the compressed domain (between the compressed source signal $F(x)$ and the not-yet-expanded received signal $F(y)$) if an "optimal" companding scheme (in the sense the authors define) is applied. Consequently, all schemes optimal for the MSE measure are also optimal for this type of non-MSE measures upon application of the optimal companding scheme. It was also shown that such an optimal compander does only depend on the distortion measure and not on the source distribution, and that it reaches the rate-distortion function asymptotically. In other words, the entropy at the output of the quantizer at a given distortion level D gets arbitrarily close to the $R(D)$ function with increasing vector dimension if an (asymptotically) optimal compander and quantizer are used. The asymptotical optimality refers to the condition of increasing vector dimension N.

The advantage of using an optimal companding scheme in audio source coding is then that

[1] $m = 0$ is a trivial case where the receiver cannot reconstruct anything and where we can only apply concealing methods.

coding with a perceptually relevant (non-MSE) locally quadratic distortion measure can be done with any MSE-based scheme without the need to pre-normalize the input signal with perceptual weights. The transmission of this side information is thereby removed, and multiple description coding can be done with an arbitrary large number of descriptions n without performance loss. Although the usage of multidimensional companding seems promising, an optimal companding scheme does, in general, not exist [38]. To work around that problem, a *suboptimal* companding scheme must be built, having an additional rate (entropy), the so-called *rate loss*, at the output of the quantizer in relation to the rate that would be theoretically achievable with the (non-existent) optimal companding scheme.

1.3 Problem Formulation

After having gotten the reader acquaintanced with the elements involved in this thesis it is now time to state the problem. This thesis focuses on developing and simulating a multidimensional companding scheme for the perceptual distortion measure developed by S. van de Par et al., described in [59], which has sinusoidal audio coding as its main target. As no optimal companding scheme exists for this distortion measure for finite vector dimension (this will be proven in Chapter 3), a suboptimal scheme is developed. Nevertheless, the scheme should have vanishing rate-loss when the vector dimension goes to infinity, i.e., the scheme should be asymptotically optimal.

The relevance of the assignment of this thesis is that, for high vector dimension, if the performance of the companding scheme is close to the optimal performance, any quantization scheme that was developed to perform well for the MSE distortion measure will also perform well for the perceptual distortion measure of [59]. This enables applications of sinusoidal audio coding to take advantage of the most state-of-the-art quantization schemes developed for the MSE, without the need for transmission of side information. The applications that would profit most from this scheme are multiple description audio coding applications which were developed for the MSE like the one in [65], where the replacement of the perceptual normalization step by multidimensional companding would make possible the usage of an arbitrarily large number of descriptions without degradation of the performance of the system.

1.4 Proposed Approach

Our proposed approach will follow the same high-level guidelines as the one in [28], which addresses a similar problem, although with a distinct distortion measure. As it was shown in [38], the condition that defines an optimal compressor is a partial differential equation system expressed in terms of a so-called *sensitivity matrix*, which collects partial derivatives of the distortion measure. The optimal expander is then the inverse of the optimal compressor.

We propose thus to start solving the problem of designing a companding scheme by calculating this sensitivity matrix for the distortion measure in question. Afterwards, we check the possibility of building an optimal companding scheme. As we will see, the answer is negative, so that a sub-optimal scheme will have to be built. We propose to build the compressor in the frequency domain: the input signal is windowed, the Discrete Fourier Transform (DFT) is applied, then the result is passed through a non-linear function and finally the inverse DFT is applied to deliver the output of the compressor. The optimality condition in terms of this non-linear function in the frequency domain is derived, leading also to a partial differential equations system. We show that a good choice for the suboptimal non-linear function is obtained by integrating some of the equations on the system. More specifically, the equation system is a matrix equation and we solve the equations on the diagonal of the matrices on the left and right-hand side of the equation.

After building this compressor, we motivate its approximation by a Taylor expansion. Indeed, the direct calculation of an integral corresponding to the diagonal elements of the equation system is computationally expensive. Moreover, the Taylor expansion will reveal the most dominant term of the compressor, enabling its intuitive interpretation. We thus execute the derivation of the Taylor expansion and then analyze the compressor and its proximity to the optimality condition, proving that optimality is achieved asymptotically, i.e., with increasing vector dimension N. By that time, the development and analysis of the compressor will be finished so that we turn on to the expander. We show that the compressor function is invertible, at least in its simplest form, achieving the guarantee that in this case the expander exists. We then develop an algorithm based on numerical methods to compute its function value upon the insertion of a given argument. The expander algorithm is further optimized in such a way that its memory usage is reduced.

The described extensive and technical part is complemented by simulations of the companding scheme, whose results corroborate the theoretically derived asymptotical optimality. Limitations of the compander are explained and discussed on basis of additional simulations.

1.5 Contributions of this Thesis

The main contribution of this work is a suboptimal companding scheme (compressor and expander) which reaches optimality asymptotically, i.e., which has a vanishing rate-loss with increasing vector dimension. The compressor consists in a frequency domain application of a non-linear function. This function multiplies the input signal component-wise by a signal-dependent gain. In its rawest form, the signal-dependent gain is an indefinite integration of the (signal dependent) square-root of the inverse of the masking threshold with respect to each component of the frequency domain signal. We derive that the dominating Taylor expansion term of the gain is exactly the integrand of it, the square-root of the inverse of the masking threshold. The compressor, in its simplest and most elegant form, thus boils down to the point-wise multiplication of the windowed input signal by the sensitivity the human ear has to it (the inverse of the masking threshold in signal dimensions). Although the compressor obtained from this simplest form of the gain is exactly equivalent to performing normalization of the input signal by perceptual weights, the proposed approach has the novelty of avoiding the transmission of the perceptual weights through the channel. Indeed, at the receiver we only have to run the inverse of the compressor (i.e., the expander) to reconstruct the signal, and the expander does not depend on the perceptual weights.

In addition to the development of a compressor, we also develop the corresponding expander (the inverse function) on basis of well justified numerical methods: we perform one iteration on a fixed point equation (a rewritten form of the equation defining the compressor) and fine-tune the result using Broyden's method [44].

An additional original contribution of the thesis is the proof that no optimal companding scheme exists, assuming certain restrictions on the condition of optimality.

Side-contributions are the delivery of the rate-distortion function for the distortion measure taken into consideration [59] at high-resolution, based on [37], and also of the rate-loss of the developed scheme, based on [38].

1.6 Thesis Paper Organization

For guiding the reader, the structure of the thesis paper is outlined as follows. We start in Chapter 2 by technically introducing the topics of locally quadratic distortion measures (Section 2.1), multidimensional companding (Section 2.2) and the (locally quadratic) perceptual distortion measure of S. van de Par et al. [59] (Section 2.3). The main content of this thesis is exposed in Chapter 3. In Section 3.1, the sensitivity matrix for this distortion measure is calculated; in Section 3.2, the rawest form of the compressor is developed; in Section 3.3, the compressor is worked out in a Taylor expansion; in Section 3.4, it is analyzed in terms of its Jacobian matrix and its asymptotic optimality is established; finally, in Section 3.5, the expander is developed on basis of numerical methods. Chapter 4 is devoted to simulations. We first calculate the rate-distortion function for the distortion measure at high resolution (Section 4.1), then the rate-loss incurred by the developed suboptimal companding scheme (Section 4.2) and finally we perform simulations based on the calculated quantities and derived results (Section 4.3). In Chapter 5, conclusions are drawn and directions for future work regarding the topic of this thesis are discussed.

Chapter 2

Technical Overview

We will start this thesis with a technical overview of the main toolboxes that it will deal with: locally quadratic distortion measures (in general) [37, 38, 51], multidimensional companding [38] and the distortion measure developed by S. van de Par et al. [59]. As the current work is developed in the framework of locally quadratic distortion measures, this type of measures is overviewed in the first place, and recent results on the rate-distortion function for them are mentioned. Following that, it will be time to explain multidimensional companding source coding. The condition for the optimality of a companding scheme is deduced using heuristical arguments (and of course referred for the detailed formal explanation), and conditions for the existence of optimal schemes are explored. A benchmark criterion for non-optimal schemes is then delivered and finally a summary of previous work on multidimensional companding is given. The third and last part of the chapter is dedicated to a locally quadratic perceptual distortion measure which will be used throughout this work. The measure is presented, re-arranged in several convenient forms and discussed.

2.1 Locally Quadratic Distortion Measures and Correspondent $R(D)$ function

In this section, we characterize locally quadratic distortion measures and deliver their rate-distortion function at high resolution on basis of the work of Linder et al. [37,38]. Locally quadratic distortion measures are measures satisfying smoothness conditions at high resolution (when the reproduction signal approaches the source signal), and which are positive definite. Their rate-distortion function at high resolution is similar to the Shannon Lower Bound [9], but with an excess term dependent on the distortion measure.

2.1.1 Locally Quadratic Distortion Measures

In [38], multidimensional companding theory was studied assuming that a *locally quadratic* distortion measure $d(x, y)$ between the source x and reproduction y was used. A locally quadratic distortion measure is characterized by being of class $C^3(\mathbb{R}^N)$ (d is 3 times continuously differentiable) with respect to y, by being strictly positive except in its absolute minimum $y = x$, where it should be 0,

$$d(x, y) \geq 0 \quad \text{with equality iff } y = x, \tag{2.1}$$

and by having a Taylor expansion with respect to y around x given by

$$\begin{aligned} d(x, y) &= d(x, x) + \nabla_y[d(x, y)]\Big|_{y=x}(y - x) + (y - x)^{\mathrm{T}} M(x)(y - x) + O(\|y - x\|^3) && (2.2)\\ &= (y - x)^{\mathrm{T}} M(x)(y - x) + O(\|y - x\|^3) && (2.3)\\ &= \left[\sum_{m,l=0}^{N-1} [M(x)]_{m,l}\,(y_m - x_m)(y_l - x_l)\right] + O(\|y - x\|^3), && (2.4)\end{aligned}$$

where ∇_y is the gradient operator with respect to the vector y, where we use the big O notation with the symbol $O(\cdot)$, where $\|\cdot\|$ denotes the l_2 norm and where $M(x)$ is defined as

$$[M(x)]_{m,l} = \frac{1}{2}\frac{\partial^2 d(x,y)}{\partial y_m \partial y_l}\Big|_{y=x}, \quad m,l = 0, 1, \ldots, N - 1, \tag{2.5}$$

i.e., it is half of the Hessian matrix of d with respect to y calculated at $y = x$, dubbed in [38] as the *sensitivity matrix.* Note that in this work, indexing of vectors and matrices is represented by $[V]_{\cdot,\cdot}$ and $v_\cdot$, respectively, and will start at 0. An alternative representation for the vector indexing is $v(\cdot)$. Furthermore, note also that the disappearance of the 0^{th} and 1^{st} order term in (2.3) comes directly from the condition (2.1). Due to the same condition, $M(x)$ is positive definite.

Equation (2.3) delivers the essence of a locally quadratic distortion measure: at high resolution ($y \to x$), the distortion between x and y can be approximated by a quadratic form applied to $y - x$.

2.1.2 $R(D)$ Function of a Locally Quadratic Distortion Measure

For studying the performance of the companding scheme to be developed, it is of theoretical interest to know what the best ever achievable performance is, independently of using a companding scheme or not. This best possible is given by the Shannon's rate-distortion function $R(D)$, which depends on the used distortion measure d and on the source X. It delivers the theoretical minimum rate per dimension at which it is possible to code the source X, given that we are coding at an expected distortion level per dimension $\mathrm{E}[d(X, Y(X))]/N$ not greater than D. This function is obtained through the minimization problem [9]

$$R(D) = \inf\{I(X,Y) : \mathrm{E}[d(X,Y)]/N \leq D\}, \tag{2.6}$$

where the infimum is taken over all possible conditional distributions of Y given X and $I(X,Y)$ denotes the mutual information between source and reproduction per dimension, i.e., being p the (joint/marginal-)pdf of its uppercased argument(s),

$$I(X,Y) = \frac{1}{N} \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)\,p(y)} \right) \mathrm{d}x\,\mathrm{d}y. \tag{2.7}$$

Due to its variational nature, the optimization problem (2.6) is in general difficult to solve, and to date its solution is only known for some special cases of sources and/or special distortion measures. Nevertheless, Linder et al. found an expression for the rate-distortion function at high resolution for an arbitrary source and a locally quadratic distortion measure (Subsection 2.1.1), provided that they fulfill certain regularity conditions [37], namely

$$\lim_{D\to 0} \left[R(D) + \frac{1}{2} \log_2(2\pi \mathrm{e} D) \right] = h(X) + \frac{1}{2N} \mathrm{E}[\log_2 \det M(X)], \tag{2.8}$$

where $h(X)$ denotes the differential entropy of the source per dimension,

$$h(X) = -\frac{1}{N} \int_{\mathbb{R}^N} p(x) \log_2 p(x)\,\mathrm{d}x. \tag{2.9}$$

In approximate terms we have thus, at high resolution,

$$R(D) \approx h(X) - \frac{1}{2} \log_2(2\pi \mathrm{e} D) + \frac{1}{2N} \mathrm{E}[\log_2 \det M(X)], \tag{2.10}$$

This function parallels the usual Shannon lower bound [9], where an additional term is added to take into account the fact that we are using a distortion measure different from the MSE; although quadratic, this distortion measure gives (in general) different importance to the contribution of each component $(y-x)_m(y-x)_l,\ m,l = 0,1,\ldots,N-1$ to the overall distortion (cf. Equation (2.4)). An alternative way to express the theoretical limit (2.10) is to use its inverse, the distortion-rate function

$$D(R) \approx \frac{1}{2\pi \mathrm{e}}\, 2^{2\left(h(X) - R + \frac{1}{2N}\mathrm{E}[\log_2 \det M(X)]\right)}. \tag{2.11}$$

From Equation (2.10), we see that (at high resolution) the rate-distortion function is only dependent on the differential entropy of the source $h(X)$ and on the average behavior of the sensitivity matrix $M(X)$, which in turn depends on the distortion measure.

2.2 Multidimensional Companding

We will now explain the basic concepts present in a multidimensional companding source coding scheme in this section, namely its basic structure and the type of quantization it uses. We then proceed explaining how a scheme is defined to be optimal, and we deliver a necessary and sufficient condition for the existence of an optimal scheme, given a constraint on the exposed condition for optimality. A benchmark criterion for sub-optimal schemes is exposed using results of [38], and finally the work done in the scientific community on multidimensional companding is overviewed.

2.2.1 Overview of Multidimensional Companding

Multidimensional companding is the simple operation that can be seen in Figure 1.2. In a multidimensional (say dimension $N > 1$) source coding scheme, instead of quantizing the source x directly, transmitting through the channel in an entropically efficient way and consulting a codebook (inverse quantization) to obtain the reconstruction y, we apply a pre-processing vector function F, the compressor, before quantizing and its inverse F^{-1} as a post-processing step, the expander, after inverse quantization. Obviously, this places the restriction on F that it must be invertible, at least locally (in a neighborhood of all function values $F(x)$). Indeed, if this local invertibility exists and the resolution is sufficiently high ($y \to x$), then we fall in the region of local invertibility of the function and we are sure that there is one and only one value of y which generates $F(y)$ (the inverse $F^{-1}\{Q^{-1}[Q(x)]\}$). In this thesis, we will deal with real-valued sources $x \in \mathbb{R}^N$, so that all signals on the arrows of Figure 1.2 are real (multidimensional) signals and all functions are real-valued and have real arguments.

2.2.2 Quantization

The type of quantization that we use in multidimensional companding is *Lattice Vector Quantization* (LVQ). In this type of quantization, the quantizer consists of a grid of points, the *lattice* Λ, which is generated by a linear combination of N-dimensional basis vectors with all possible integer coefficients,

$$\Lambda = \left\{ v \in \mathbb{R}^N : v = \sum_{i=0}^{N-1} n_i v_i,\ \forall n_i \in \mathbb{Z} \right\}, \tag{2.12}$$

being thus the lattice defined by the linear independent vectors $v_0, v_1, \ldots, v_{N-1} \in \mathbb{R}^N$. To make quantization with different step sizes possible (i.e., with different entropy at the output of the quantizer), we introduce the scaled lattice $s\Lambda$, defined by

$$s\Lambda = \left\{ sv : v \in \Lambda,\, s \in \mathbb{R}^+ \right\}. \tag{2.13}$$

s is thus a generalization of the quantizer step size in LVQ. The quantizer which corresponds to the lattice $s\Lambda$, which we will denote by $Q_{s\Lambda}(\cdot)$, then works by approximating a certain vector input z by the nearest element of the lattice $s\Lambda$, i.e.,

$$Q_{s\Lambda}(z) = \{v \in s\Lambda : \|v - z\| \leq \|v' - z\|,\ \forall v' \in s\Lambda\}\,. \tag{2.14}$$

We call the region of all points z mapped into a certain quantized vector $v = Q_{s\Lambda}(z)$ a *Voronoi cell.* For $v = 0$, this cell is called the *basic cell* of the lattice. In addition, a relevant quantity associated to a lattice is its *normalized moment of inertia* or *normalized second moment*, given by

$$G_\Lambda = \frac{1}{N\mathrm{Vol}(\Lambda)^{1+2/N}} \int_{\substack{\text{basic}\\\text{cell}}} \|x\|^2 \,\mathrm{d}x, \tag{2.15}$$

where $\mathrm{Vol}(\cdot)$ denotes the volume of a Voronoi cell of Λ. This quantity is scale independent ($G_{s\Lambda} = G_\Lambda$) and is a benchmark value of the lattice vector quantizer, since the average MSE distortion is proportional to it. The lowest possible value of G_Λ is $1/(2\pi\mathrm{e})$, the normalized moment of an infinite-dimensional sphere. The sphere-packing loss for a certain lattice vector quantizer with lattice Λ is then defined as the additional rate derived from not using the infinite-dimensional sphere,

$$\mathrm{SL}_R = \frac{1}{2}\log_2(2\pi\mathrm{e}G_\Lambda) \tag{2.16}$$

or as the additional distortion (in dB)

$$\mathrm{SL}_D = 10\log_{10}(2\pi\mathrm{e}G_\Lambda). \tag{2.17}$$

For each possible output of the quantizer, a symbol is assigned, being the resultant symbols coded by the succeeding entropy coder. At the receiver, the entropy decoded symbols are converted back to the reproduction points (2.14) through consultation of a table with such a mapping, which we call the *codebook.* Note that in Figure 1.2, we denoted the chain of the quantizer and symbol assigner by $Q(\cdot)$ and the codebook by $Q^{-1}(\cdot)$. From this section on, we will merge both steps into $Q_{s\Lambda}(\cdot)$ for notational simplicity. Lattice vector quantization has the advantage of being conceptually simple, and that a wide range of literature in high resolution quantization theory is based on it, since quantization of this type (at high resolution) when using the MSE as distortion criterion is thought to be optimal at high resolution, independently of the source distribution [20,39]. We refer the reader to [21,23,24] for an overview on vector and high resolution quantization. We will not dive deeply into the entropy coding / channel transmission block of Figure 1.2 in this work, assuming simply that there are high-efficiency entropy coders that can produce an average code length very close to (but greater than or equal to) the entropy of the quantizer Q, that the channel has a capacity greater or equal than that entropy, and that there are high-efficiency channel coders as well, which can code at a rate arbitrarily close to the channel capacity with an arbitrarily low error rate. In other words, we assume that there are entropy and channel coders that can approach very well the theoretical limits of Shannon's source and channel coding theorems [14].

2.2.3 Optimality

It was derived in [38] that a multidimensional companding scheme is optimal when the Jacobian matrix

$$[F'(x)]_{m,l} = \frac{\partial F_m}{\partial x_l}(x), \quad m,l = 0, 1, \ldots, N-1 \tag{2.18}$$

of the compressor satisfies[1]

$$M(X) = F'(X)^{\mathrm{T}} F'(X) \tag{2.19}$$

almost everywhere, where $M(X)$ is the sensitivity matrix of Equation (2.5) and where we denote random variables by uppercase letters and their realizations by the lowercase correspondents. It was also shown that for an optimal scheme, at high resolution, the distortion measure $d(x, y)$ gets transformed into the squared distortion measure between $F(x)$ and $F(y)$. To get an intuition on the optimality condition (2.19), we will show this property here using heuristical arguments. Consider the case of high resolution, $y \to x$. In that case, we can state that the 2^{nd} order term of the Taylor expansion (2.3) is dominant, so that we have approximately

$$d(x, y) \approx (y - x)^{\mathrm{T}} M(x)(y - x). \tag{2.20}$$

On the other hand, for a differentiable compressor F, also for $y \to x$, we can approximate the finite difference $F(y) - F(x)$ by

$$F(y) - F(x) \approx F'(x)(y - x). \tag{2.21}$$

This is known as the secant equation, obtained by a simple 1^{st} order Taylor expansion of F on y around x. The squared norm of this difference is then approximately

$$\|F(y) - F(x)\|^2 \approx \|F'(x)(y - x)\|^2 = (y - x)^{\mathrm{T}} F'(x)^{\mathrm{T}} F'(x)(y - x), \tag{2.22}$$

so that when (2.19) is valid, (2.20) results, and consequently

$$d(x, y) \approx \|F(y) - F(x)\|^2. \tag{2.23}$$

As LVQ achieves the rate-distortion bound at high resolution asymptotically for the MSE distortion measure for any source distribution, it is then intuitive that the rate-distortion bound is achieved

[1] In the original paper there was a multiplying constant c which we set here to 1 for simplicity. No loss of generality occurs.

with the optimal compander (compressor and expander) satisfying (2.19) for the locally quadratic distortion measure d. Indeed, denoting by $H(Q_{D,F})$ the entropy per dimension at the output of the lattice vector quantizer Q upon the usage of a companding scheme with compressor F and by $R(D)$ the Shannon's rate-distortion function (with a dimension-normalized rate), in [38] it is also proven that

$$\lim_{D\to 0}\left[H(Q_{D,F}) - R(D)\right] = \frac{1}{2}\log_2(2\pi e\, G_\Lambda), \tag{2.24}$$

where G_Λ is the normalized moment of inertia of the lattice. As the normalized moment of inertia of the optimal Lattice Vector Quantizer converges to $1/(2\pi e)$ when $N \to \infty$ (the normalized moment of an infinite-dimensional sphere) [63], the performance of an optimal companding scheme approaches the best possible performance when $N \to \infty$, given by the $R(D)$ function. For a more detailed discussion, see [38].

2.2.4 Existence

Although rate-distortion optimality can be achieved at high resolution when an optimal multidimensional companding scheme is used, such a scheme does not always exist [20]. An intuitive argumentation for this fact is that there are more equations in (2.19) (N^2 equations, corresponding to the components of the matrices) than unknowns (N unknowns, corresponding to the N coordinate functions). To see this in more detail, note that due to the property (2.1), the sensitivity matrix (2.5) is positive definite, so that there exists a matrix F' satisfying (2.19) [29]. We call any matrix satisfying Equation (2.19) a square-root of M, denoted here by $\sqrt{M}$. An equivalent condition to (2.19) is then

$$F'(x) = \sqrt{M(x)}, \tag{2.25}$$

for some square-root of M. Assuming that the second partial derivatives of F exist and are continuous, Schwarz' theorem has to hold, i.e.,

$$\frac{\partial^2 F_m}{\partial x_k \partial x_l}(x) = \frac{\partial^2 F_m}{\partial x_l \partial x_k}(x), \quad \forall m,l,k = 0,1,\ldots,N-1. \tag{2.26}$$

In terms of Equation (2.25), we must have

$$\frac{\partial[\sqrt{M(x)}]_{m,l}}{\partial x_k} = \frac{\partial[\sqrt{M(x)}]_{m,k}}{\partial x_l}, \quad \forall m,l,k = 0,1,\ldots,N-1. \tag{2.27}$$

This equation poses a restriction on the square-root of the sensitivity matrix (dependent on the distortion measure), so that in general, for a given square-root of the sensitivity matrix, an optimal companding scheme does not exist. Reciprocally, it is easy so see that if (2.27) is fulfilled, then the optimal compressor exists and that it is given by the anti-derivative

$$F_m(x) = \int [\sqrt{M(x)}]_{m,k}\,\mathrm{d}x_k, \quad m = 0,1,\ldots,N-1, \tag{2.28}$$

where $k \in \{0, 1, \ldots, N-1\}$ can be chosen arbitrarily. Indeed if we choose such a compressor we get

$$\frac{\partial F_m}{\partial x_l}(x) = \int \frac{\partial [\sqrt{M(x)}]_{m,k}}{\partial x_l}\,\mathrm{d}x_k = \int \frac{\partial [\sqrt{M(x)}]_{m,l}}{\partial x_k}\,\mathrm{d}x_k = [\sqrt{M(x)}]_{m,l}. \tag{2.29}$$

Note that the square-root of a matrix is not unique, and that this analysis was made for a given square-root of the sensitivity matrix. It is possible that for a given square-root no solution of (2.25) exists (equivalent to (2.27) being satisfied, upon assumption of regularity conditions for F) but for another square-root a solution does exist. As an example, you can take $N = 2$ with $x = [x_0, x_1]^{\mathrm{T}} \in \mathbb{R}^2$ and the magnitude normalized distortion measure

$$d_{\mathrm{MN}}(x, y) = \sum_{n=0}^{1} \frac{(y_n - x_n)^2}{x_n^2}, \tag{2.30}$$

where the sensitivity matrix (2.5) is [28]

$$M_{\mathrm{MN}}(x) = \begin{bmatrix} \frac{1}{x_0^2} & 0 \\ 0 & \frac{1}{x_1^2} \end{bmatrix} \tag{2.31}$$

and two possible square-roots are

$$F'_{\mathrm{MN}}(x) = \begin{bmatrix} \frac{1}{x_0} & 0 \\ 0 & \frac{1}{x_1} \end{bmatrix}, \qquad F'_{\mathrm{MN},2}(x) = \begin{bmatrix} \frac{\sin x_0}{x_0} & -\frac{\cos x_0}{x_1} \\ \frac{\cos x_0}{x_0} & \frac{\sin x_0}{x_1} \end{bmatrix}. \tag{2.32}$$

It is easy to see through (2.27) that $[F_{\mathrm{MN}}(x)]_n = \ln |x_n|,\ n = 0, 1$ exists whereas $F_{\mathrm{MN},2}$ does not. Under which conditions may then such a thing happen?

To analyze this situation we need to know the form of all square-roots of the sensitivity matrix. Through simple calculations, we see that any other solution $F'_2(x)$ of (2.19) satisfies

$$F'_2(x) = F'_2(x)^{-\mathrm{T}} M(x) = F'_2(x)^{-\mathrm{T}} F'(x)^{\mathrm{T}} F'(x) = U(x) F'(x), \tag{2.33}$$

where we define

$$U(x) \overset{\mathrm{def}}{=} F'_2(x)^{-\mathrm{T}} F'(x)^{\mathrm{T}}, \tag{2.34}$$

a matrix which satisfies (I is the identity matrix)

$$U(x)^{\mathrm{T}} U(x) = \left(F'_2(x)^{-\mathrm{T}} F'(x)^{\mathrm{T}}\right)^{\mathrm{T}} \left(F'_2(x)^{-\mathrm{T}} F'(x)^{\mathrm{T}}\right) = F'(x) F'_2(x)^{-1} F'_2(x)^{-\mathrm{T}} F'(x)^{\mathrm{T}} \tag{2.35}$$

$$= F'(x) M(x)^{-1} F'(x)^{\mathrm{T}} = F'(x) F'(x)^{-1} F'(x)^{-\mathrm{T}} F'(x)^{\mathrm{T}} = \mathrm{I}, \tag{2.36}$$

i.e., which is orthogonal (you can do the same calculations for $U(x)U(x)^{\mathrm{T}} = \mathrm{I}$). Note that both $F'(x)$ and $F'_2(x)$ must be invertible due to the fact that $M(x)$ is strictly positive definite. We have thus shown that all possible solutions are given in terms of a certain square-root $F'(x) = \sqrt{M(x)}$

by the left multiplication by an orthogonal matrix $U(x)$.

Imagine now that U is independent of x. We state that $F'(x) = \sqrt{M(x)}$ has a solution $F(x)$ if and only if $F_2'(x) = U\sqrt{M(x)}$ has a solution $F_2(x)$ as well, for any orthogonal matrix U independent of x. The proof is as follows. If F exists, then we define $F_2(x) \overset{\text{def}}{=} UF(x)$, and using simple differentiation rules [5], F_2 has a derivative $F_2'(x) = UF'(x)$. Equation (2.19) is thus satisfied for F_2 due to the orthogonality of U, and we have thus proven that an F_2 exists. The reciprocal relation is obtained following a similar procedure for $F(x) \overset{\text{def}}{=} U^{-1}F_2(x)$. Note that U is invertible due to being orthogonal.

Summing up the results of the previous paragraph, all square-roots of the sensitivity matrix are equal up to the left multiplication by an orthogonal matrix $U(x)$ and for U not dependent on x, an optimal companding scheme with compander $\{F, F^{-1}\}$ exists if and only if an optimal scheme does also exist for any square-root of the form $U\sqrt{M(x)}$, with U orthogonal. Although an orthogonal freedom of $F'(x)$ exists, in this thesis we will only deal with the case $U(x) = I$. We are thus assured that if for the calculated square-root $\sqrt{M(x)}$ an optimal companding scheme does not exist, then for all other square-roots $U\sqrt{M(x)}$, such a scheme does not exist as well. The case $U(x)$ dependent on x will not be studied in this thesis.

2.2.5 Rate-loss

When an optimal companding scheme does not exist, it is convenient to quantify the penalty in performance when a sub-optimal (non-optimal) scheme is used, with respect to the optimal one. Linder et al. found an expression for the rate-distortion performance of an arbitrary companding scheme $\{\tilde{F}, \tilde{F}^{-1}\}$ at high resolution, assuming weak conditions on the distribution of the source x and the usage of a locally quadratic distortion measure [38]. Their result was

$$\lim_{D\to 0}\left[H(Q_{D,\tilde{F}}) + \frac{1}{2}\log_2(D)\right] = h(X) + \frac{1}{N}\mathrm{E}\log_2|\det\tilde{F}'(X)| + \frac{1}{2}\log_2\left\{G_\Lambda\,\mathrm{E}\ \mathrm{tr}\left[\tilde{M}(X)^{-1}M(X)\right]\right\}, \tag{2.37}$$

where $H(Q_{D,\tilde{F}})$ denotes the entropy per dimension at the output of the quantizing block when coding with a multidimensional companding scheme with a compressor $\tilde{F}$ and at an average distortion per dimension D, where $h(X)$ denotes the differential entropy of the source X per dimension, where tr denotes the trace operator and where

$$\tilde{M}(X) = \tilde{F}'(X)^{\mathrm{T}}\tilde{F}'(X). \tag{2.38}$$

Equation (2.37) can be used to create a benchmark criterion for a sub-optimal compressor $\tilde{F}$. If we denote the optimal compressor by F, then the additional rate at the quantizer block of Figure 1.2 when using $\tilde{F}$ with respect to the minimum possible rate, which is achieved when we use the optimal compressor F, when coding at the same distortion level D (at high resolution) is given by

$$H(Q_{D,\tilde{F}}) - H(Q_{D,F}) \approx \frac{1}{2N} \mathrm{E} \log_2 \left(\frac{\det \tilde{M}(X)}{\det M(X)} \right) + \frac{1}{2} \log_2 \left[\mathrm{E}\ \mathrm{tr} \left(\frac{\tilde{M}(X)^{-1} M(X)}{N} \right) \right]. \tag{2.39}$$

This quantity is baptized as the *rate-loss* when using a suboptimal compander $\{\tilde{F}, \tilde{F}^{-1}\}$. Of course we can do the same in terms of the distortions $D(Q_{H,\tilde{F}})$ and $D(Q_{H,F})$ when coding at the same rate H, obtaining

$$\frac{D(Q_{H,\tilde{F}})}{D(Q_{H,F})} \approx 2^{2(H(Q_{D,\tilde{F}}) - H(Q_{D,F}))}, \tag{2.40}$$

or in dB,

$$\left(\frac{D(Q_{H,\tilde{F}})}{D(Q_{H,F})} \right)_{\mathrm{dB}} \approx 20 \log_{10}(2) \left[H(Q_{D,\tilde{F}}) - H(Q_{D,F}) \right]. \tag{2.41}$$

Equation (2.41) should be understood as the perceptual distortion power increase that we get in dB when coding with the sub-optimal compressor $\tilde{F}$, where $D = \mathrm{E}[d(X,Y)]$ comes from a perceptual distortion measure (different from the MSE).

The lower the equations (2.39), (2.40) and (2.41) are, the closer we are to the optimal compressor F, so that these equations are a way of analyzing the performance of a sub-optimal compressor. A nice property for a sub-optimal compressor is, for example, that these quantified losses disappear with increasing vector dimension ($N \to \infty$). If that happens, we say that the sub-optimal compressor is asymptotically optimal, and, if we use a sufficiently large N in practice, we will operate closely to the optimum. This asymptotical optimality is the design objective for the companding scheme to be developed in this thesis.

2.2.6 Previous Work

A companding quantization scheme was first considered by Bennett [7] for the scalar case ($N = 1$), where the source was passed through a function, quantized with a uniform scalar quantizer and then passed through the inverse function. Bennett proved that any non-uniform scalar quantization scheme could be implemented in such a way. Also for $N = 1$, Panter and Dite derived an expression for the MSE of an MSE-optimal non-uniform quantization scheme at high resolution [47]. Zador [63] generalized [47] to the multidimensional case ($N > 1$) and for an r^{th} power distortion measure $d(x,y) = d(x,y) = \|y - x\|^r$ (where the norm is l_2).

Gersho [20] unified the work that had been done so far in non-uniform quantization. Using heuristical arguments, he derived an expression for the r^{th} power distortion of a quantizer at high resolution with a certain quantizer point density, which ended up in being given in terms of an integral involving the probability density function of the source and the quantizer point density. This expression was a generalization of Bennett's work for multiple dimensions and for the r^{th} power distortion: the latter work was done for the scalar case and for the MSE. Gersho also re-derived expressions for the average r^{th} power distortion of the optimal vector quantizer of

Zador's work, given in terms of the normalized r^{th} moment of inertia $G_{\Lambda,r}$, and delivered lower and upper bounds for G. Finally, Gersho introduced the concept of multidimensional companding, and he noted that in general an optimal compander does not exist. In the case of the r^{th} power distortion measure, an optimal compressor would have to be conformal [20].

After the breakthrough of Gersho, Yamada [62] generalized Gersho's results to a more general difference distortion measure, namely an arbitrary semi-norm of the difference of reproduction and source. Bucklew [10] analyzed the performance of a multidimensional companding scheme in terms of its MSE and proved formally that an optimal compressor must be conformal. Nevertheless, he showed with an example that even with a suboptimal companding scheme, we can achieve optimality asymptotically (with increasing vector dimension N). In [11], he explored which types of probability density functions of the source enable an optimal compander.

More recently, Moo [42] developed an asymptotically optimal compressor function for the case of memoryless stationary sources. The function consisted in the independent application of a compressor to each component, dubbed in literature as the *Cartesian* compressor. Simon [55] quantified the loss introduced by a sub-optimal compander as the quotient of normalized second moments before and after the expansion operation, applying his definition to the example of a spherically symmetric (radial) compander. Linder et al. [38] developed a rigorous theory on multidimensional companding for locally quadratic distortion measures at high resolution. Also in this case, a multidimensional companding scheme does in general not exist, but if it does exist, it does not depend on the distribution of the source and it can approach the rate-distortion bound arbitrarily close. Samuelsson [53] studied the Cartesian and radial multidimensional companding schemes of [42, 55] for quantizing independent and identically distributed Gaussian sources using shaped lattice quantization ($Q^{(\text{shape})}(v) = RQ(R^{-1}v)$, where R is a signal independent matrix) to correct for linear sub-optimalities of the compressor. The problem of finding the best companding scheme through all the existent ones remains unsolved [53].

Regarding applications of multidimensional companding, piecewise linear [30], Cartesian [57] and logarithmic [28] companders were applied in image, speech and audio coding, respectively.

2.3 A Perceptually Relevant Distortion Measure

In this section the distortion measure that will be used throughout this work [59] is presented. It is a locally quadratic distortion measure used in sinusoidal audio coding with a simple closed-form mathematical expression. The distortion measure is displayed in different formulations, each one characterizing the same measure from a different point of view. Its behavior with increasing vector size N is then studied. In the end of the section, some considerations are done regarding the related work of third parties, regarding the choice of this distortion measure and regarding its advantage with respect to other measures. Finally, the aspect of perceptual weighting, explained in the introduction, is revisited on hand of the particular example of this distortion measure.

2.3.1 Distortion Measure

In [59], S. van de Par et al. defined an auditory, perceptually relevant distortion measure to be used in the context of sinusoidal audio coding, which delivers the distortion detectability of an N-dimensional signal x and its reproduction y as a weighted mean square error of the windowed signals in the frequency domain. More precisely, the distortion measure between x and y is defined as

$$d(x,y) = \sum_{f=0}^{L-1} \hat{a}^2(x,f)\, |\widehat{yw}(f) - \widehat{xw}(f)|^2, \tag{2.42}$$

where the juxtaposition of two vectors denotes the point-wise product between them, the $\hat{}$ operator denotes the L-point unnormalized Discrete Fourier Transform (DFT) in which the input signal is zero-padded up to size L, i.e., for any N-dimensional signal v

$$\hat{v}(f) = \sum_{n=0}^{N-1} v(n)\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{L}fn}, \quad f = 0, 1, \ldots, L-1 \tag{2.43}$$

with $L \geq N$, where w is an N-size frequency selective window with $w(n) > 0$, $\forall n$ and where $\hat{a}^2(x,f)$ is selected to be the (signal dependent) inverse of the masking threshold at frequency $f/L \cdot f_s$ (f_s denotes the working sample rate), given by

$$\hat{a}^2(x,f) = Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{h}_i(f)|^2}{\sum_{f'=0}^{L-1} |\hat{h}_i(f')|^2\, |\widehat{xw}(f')|^2 + c_2}, \quad f = 0, 1, \ldots, L-1. \tag{2.44}$$

In this last equation, $c_1 > 0$ and $c_2 > 0$ are calibration constants, being c_1 independent[2] of N and $h_i(n)$ is the (L-point) impulse response of the cascade of the filter simulating the behavior of the outer- and middle ear with the i^{th} gamma-tone filter of the filter-bank of size P, simulating the band-pass characteristic of the basilar membrane of the cochlea [59]. These filters h_i are assumed to be absolutely summable.

Equation (2.42) can be rewritten using matrix notation in terms of norms of vectors, and that notation will simplify the work done in this thesis. We first note that the Parseval's relation states that for any $v \in \mathbb{R}^N$ we can convert its norm in the frequency domain to the respective one in the time domain by

$$\|\hat{v}\|^2 = \sum_{f=0}^{L-1} |\hat{v}(f)|^2 = L\|v\|^2 = L \sum_{n=0}^{N-1} v^2(n). \tag{2.45}$$

[2] The case $N/f_s > 300$ ms, where the distortion measure ceases to be proportional to N, is not considered here due to the lack of stationarity of typical audio signals in that time-range.

Furthermore, let v_0 denote the zero-padded signal v, i.e.,

$$v_0(n) = \begin{cases} v(n), & n = 0, 1, \ldots, N-1 \\ 0, & n = N, N+1, \ldots, L-1. \end{cases} \tag{2.46}$$

Define also Λ_{v_0} as the (L-by-L) diagonal matrix with its elements equal to the signal v_0. Finally, denote by H_i the L-by-L circulant convolution matrix of the filter h_i, obtained by placing h_i in the first column of H_i and building the next column by circularly shifting the previous one by one unit downwards, i.e.,

$$H_i = \begin{bmatrix} h(0) & h(L-1) & h(L-2) & \cdots & h(2) & h(1) \\ h(1) & h(0) & h(L-1) & \cdots & h(3) & h(2) \\ h(2) & h(1) & h(0) & \cdots & h(4) & h(3) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ h(L-2) & h(L-3) & h(L-4) & \cdots & h(0) & h(L-1) \\ h(L-1) & h(L-2) & h(L-3) & \cdots & h(1) & h(0) \end{bmatrix}. \tag{2.47}$$

Using the Parseval's relation and these conventions, the distortion measure d can be written as

$$d(x,y) = Nc_1 \sum_{i=0}^{P-1} \frac{\sum_{f=0}^{L-1} |\hat{h}_i(f)|^2 \, |\widehat{yw}(f) - \widehat{xw}(f)|^2}{\sum_{f=0}^{L-1} |\hat{h}_i(f)|^2 \, |\widehat{xw}(f)|^2 + c_2} \tag{2.48}$$

$$= Nc_1 \sum_{i=0}^{P-1} \frac{\|H_i \Lambda_{w_0}(y_0 - x_0)\|^2}{\|H_i \Lambda_{w_0} x_0\|^2 + c_2/L} \tag{2.49}$$

In the form (2.49), the distortion measure can be interpreted as the accumulation of the filtered windowed distortions $y - x$ normalized by the filtered windowed excitation signal x, for all band-pass filters h_i (up to calibration constants). Indeed, denoting by $\circledast$ the circulant convolution operator, we can rewrite Equation (2.49) as

$$d(x,y) = Nc_1 \sum_{i=0}^{P-1} \frac{\|h_i \circledast [(y-x)w]\|^2}{\|h_i \circledast [xw]\|^2 + c_2/L}, \tag{2.50}$$

where again juxtaposition is used for point-wise multiplication.

2.3.2 Asymptotic Expression ($N \to \infty$) and Behavior

To enable the analysis of the distortion measure of Subsection 2.3.1 with increasing vector dimension N, it is convenient to reformulate it. Rewrite the distortion measure so that it uses squared magnitudes normalized by NL and a new calibration constant c_2', i.e., set

$$d(x,y) = \sum_{f=0}^{L-1} \frac{|\widehat{yw}(f) - \widehat{xw}(f)|^2}{NL} \left(Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{h}_i(f)|^2}{\sum_{f'=0}^{L-1} |\hat{h}_i(f')|^2 \frac{|\widehat{xw}(f')|^2}{NL} + c_2'} \right) \tag{2.51}$$

$$= Nc_1 \sum_{i=0}^{P-1} \frac{\sum_{f=0}^{L-1} |\hat{h}_i(f)|^2 \, |\widehat{yw}(f) - \widehat{xw}(f)|^2}{\sum_{f=0}^{L-1} |\hat{h}_i(f)|^2 \, |\widehat{xw}(f)|^2 + NL\, c_2'}. \tag{2.52}$$

The new calibration constant was introduced to make this redefinition consistent with the old definition of equations (2.42) and (2.44). Indeed, by comparison with (2.48), we get the equivalence

$$c_2 = NL\, c_2'. \tag{2.53}$$

Introduce also a new version of the inverse of the masking threshold $\hat{a}'^2(f)$ using the same normalization:

$$\hat{a}'^2(x,f) = Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{h}_i(f)|^2}{\sum_{f'=0}^{L-1} |\hat{h}_i(f')|^2 \frac{|\widehat{xw}(f')|^2}{NL} + c_2'} \tag{2.54}$$

$$= NL \left(Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{h}_i(f)|^2}{\sum_{f'=0}^{L-1} |\hat{h}_i(f')|^2 \, |\widehat{xw}(f')|^2 + NL\, c_2'} \right). \tag{2.55}$$

Again by comparison, this new function is given in terms of the old one by

$$\hat{a}'^2(x,f) = NL\, \hat{a}^2(x,f). \tag{2.56}$$

Using the Parseval's relation again, the distortion measure d and the new variant of the inverse of the masking threshold $\hat{a}'^2$ can be rewritten in terms of norms of vectors:

$$d(x,y) = Nc_1 \sum_{i=0}^{P-1} \frac{\frac{1}{N}\|H_i \Lambda_{w_0}(y_0 - x_0)\|^2}{\frac{1}{N}\|H_i \Lambda_{w_0} x_0\|^2 + c_2'}; \tag{2.57}$$

$$\hat{a}'^2(x,f) = Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{h}_i(f)|^2}{\frac{1}{N}\|H_i \Lambda_{w_0} x_0\|^2 + c_2'}. \tag{2.58}$$

With these new expressions for d and $\hat{a}'^2$, the asymptotic analysis of these two quantities becomes more simple. Indeed, assuming x wide-sense stationary[3] and ergodic, for large N, the quantities $\|H_i \Lambda_{w_0} x_0\|^2/N$ and $\|H_i \Lambda_{w_0}(y_0 - x_0)\|^2/N$ can be seen as estimates for the (time-independent) power of the filtered input and error signals, respectively. Note that for $N \to \infty$ the effect of the window disappears, so that for large N we have indeed a good power estimate. The correspondent estimators are given by the statistical average of the squared signals $H_i \Lambda_{w_0} x_0$ and $H_i \Lambda_{w_0}(y_0 - x_0)$, respectively, so that they are consistent (i.e., their variance decreases with N).

[3] For the values of N we typically deal with, e.g., $N = 1024$ at a sample frequency of 48 kHz (21,3 milliseconds), x can be assumed to be both wide-sense stationary and large enough so that the asymptotic results which are going to be presented are good approximations.

We assume additionally that c_1 and c'_2 are independent of N. As in [59], the contribution of the temporal integration time of the human auditory system is considered entirely on the proportionality constant N on the left of c_1 (we do not treat the case $N/f_s > 300$ ms). As to c'_2, it can be considered independent of N due to the fact that it is summed side-by-side with the estimated power of the input signal. It can be thus interpreted as the internal noise [59] power in the human hearing system, which independent of N.

As a result of the previous considerations, the distortion d and the inverse of the masking threshold $\hat{a}'^2$ can be approximated, for large N (but not so large that N/f_s exceeds 300 ms), by

$$d(x,y) \approx Nc_1 \sum_{i=0}^{P-1} \frac{P_{[(Y-X)*h_i]}}{P_{[X*h_i]} + c'_2} \quad \text{and} \tag{2.59}$$

$$\hat{a}'^2(x,f) \approx Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{h}_i(f)|^2}{P_{[X*h_i]} + c'_2}, \tag{2.60}$$

respectively. In these expressions, $P_{[V]} = \mathrm{E}[V_n^2]$ denotes the power of the random variable V, and $*$ denotes the linear convolution. Note that asymptotically, the linear and circulant convolutions produce the same results. Furthermore, note that both quantities (2.59) and (2.60) are proportional to N, since the expressions on the right of N are independent of it. Accordingly, d/N and $\hat{a}'^2/N$ converge for $N \to \infty$.

2.3.3 Considerations about the Distortion Measure

It is of interest to note that the distortion measure (2.42), (2.50) can be rewritten in the form of the measure studied in [28],

$$d(x,y) = \sum_{i=0}^{P-1} \frac{\|y_i - x_i\|^2}{\|x_i\|^2} \tag{2.61}$$

doing the substitutions $x_i = h_i \circledast [xw]$ and $y_i = h_i \circledast [yw]$ if we neglect the constants Nc_1 and c_2. Nevertheless, in that work, x_i (and y_i) were considered to be separate signal inputs (and reproductions) for each i, which are to be treated separately. If we were to apply the work of [28] directly to this distortion measure, the signals that we would have to transmit would thus have a dimension NP instead of N, which would be an unnecessary waste of bit-rate, since all x_i depend on the same x through h_i. Of course, due to the band-pass filtering, each x_i represents less information than the original signal x, and the application of multi-rate signal processing techniques [60] could eventually be possible to decimate the signals x_i so that in the end, we again have a signal $x_1, x_2, \ldots, x_P$ of total size N. Such techniques were not researched in this thesis.

Furthermore, it is important to mention that several other distortion criteria have been developed for audio coding, such as the noise-to-mask ratio obtained from a masking threshold based on spreading functions [2, 31], a modified version of it [41], the mean-square error of Dau's internal representation [15, 48], or a generalization of S. van de Par's measure [61], which upon the substitution of a parameter degenerated into that measure, into a measure modeling the overall *loudness* [64] of the distortion signal, or into a frequency-weighted MSE with constant weights, not taking masking effects into account. Although these distortion measures can, in certain cases, outperform the distortion measure developed by S. van de Par, they are either defined algorithmically, or, although mathematically defined, they are more complex in nature. The choice of S. van de Par's distortion measure for this work is thus motivated by the fact that it is a measure which lies on the realm of the mathematically tractable locally quadratic distortion measures (it is easy to see that indeed the measure is locally quadratic) and that it delivers proven results in sinusoidal audio coding.

Finally, in accordance to what was explained in the introduction (Chapter 1), we would like to point out that this distortion measure can be transformed into a mean-square-error (MSE) by normalization of the input signal. Indeed, looking at the form (2.42) of the distortion measure, we can define

$$\hat{x}'(f) = \hat{a}(x,f)\,\widehat{xw}(f) \qquad\qquad \hat{y}'(f) = \hat{a}(x,f)\,\widehat{yw}(f) \tag{2.62}$$

and work on the normalized domain x', y', i.e., source-code and transmit x' and recover y from y'. Nevertheless, if we use the normalization directly, the square-root of the inverse of the masking threshold $\hat{a}$ has to be transmitted to the receiver to perform the inverse normalization

$$\widehat{yw}(f) = \frac{\hat{y}'(f)}{\hat{a}(x,f)}, \tag{2.63}$$

with the consequence of an intolerable overhead in certain conditions, as explained in the introduction.

Chapter 3

A Suboptimal Companding Scheme

As explained in the introduction (Chapter 1), this thesis paper will concentrate on developing a suboptimal multidimensional companding scheme (cf. Section 2.2) for the distortion measure presented in [59], developed for sinusoidal audio coding (Section 2.3). In this chapter, we start deriving the sensitivity matrix (2.5) for the desired distortion measure, Equation (2.42). Afterwards, the optimality condition (2.19) is worked out more deeply, and we prove from it that no optimal companding scheme exists for finite vector dimension, at least in the scope studied in this work. We then derive a non-optimal compressor based on the same condition, simplify it for numerical speed-up, and then analyze it. In this step we calculate its actual (non-ideal) Jacobian Matrix (Equation (2.18) for the non-optimal compressor), and using that result we analyze the compressor in terms of its behavior with respect to the optimal one with increasing vector dimension. We show that, asymptotically, the compressor is optimal, making the rate-loss (2.39) vanish. Finally, we end this chapter building the correspondent expander based on numerical methods.

3.1 Sensitivity Matrix

In this section, the sensitivity matrix (2.5), function of the distortion measure, is derived and two different cases of the relation between matrix size and DFT-size are studied. In the first case, the sensitivity matrix is decomposed in terms of its eigenvalue decomposition and in the second case it is approximated by a matrix easily decomposable in terms of that decomposition. The importance of this derivation of the sensitivity matrix comes from the fact that it defines the conditions for the optimality of a companding scheme (cf. subsections 2.2.3 and 2.2.4).

As explained in Subsection 2.2.3, if an optimal compressor $F(x)$ exists, its derivative (Jacobian matrix) $F'(x)$ is given in terms of the sensitivity matrix $M(x)$ (2.5), which in turn depends on the distortion measure (2.42). In the first place, we note that (2.42) is a locally quadratic measure; it is easy to see that $d(x, y) \geq 0$ with equality iff $y = x$, and that it is a composition of analytical functions in y (in $\mathbb{R}^N$), thus also functions of class C^3 in y. Having justified the existence of $M(x)$, we now wish to calculate it explicitly. The expansion of d in the form (2.49) using elementary algebra ($^{\mathrm{T}}$ denotes transposition) yields

$$d(x,y) = Nc_1 \sum_{i=0}^{P-1} \frac{y_0^{\mathrm{T}} \Lambda_{w_0} H_i^{\mathrm{T}} H_i \Lambda_{w_0} y_0 - 2x_0^{\mathrm{T}} \Lambda_{w_0} H_i^{\mathrm{T}} H_i \Lambda_{w_0} y_0 + x_0^{\mathrm{T}} \Lambda_{w_0} H_i^{\mathrm{T}} H_i \Lambda_{w_0} x_0}{\|H_i \Lambda_{w_0} x_0\|^2 + c_2/L}. \tag{3.1}$$

Differentiating with respect to y_l and y_m, we obtain[1]

$$\frac{\partial\, d(x,y)}{\partial y_m} = Nc_1 \sum_{i=0}^{P-1} \frac{2[\Lambda_{w_0} H_i^{\mathrm{T}} H_i \Lambda_{w_0} (y_0 - x_0)]_m}{\|H_i \Lambda_{w_0} x_0\|^2 + c_2/L}, \tag{3.2}$$

$$\frac{\partial^2 d(x,y)}{\partial y_l \partial y_m} = Nc_1 \sum_{i=0}^{P-1} \frac{2[\Lambda_{w_0} H_i^{\mathrm{T}} H_i \Lambda_{w_0}]_{m,l}}{\|H_i \Lambda_{w_0} x_0\|^2 + c_2/L} \tag{3.3}$$

$$[M(x)]_{m,l} = \frac{1}{2} \frac{\partial^2 d(x,y)}{\partial y_m \partial y_l}\Big|_{y=x} = Nc_1 \sum_{i=0}^{P-1} \frac{[\Lambda_{w_0} H_i^{\mathrm{T}} H_i \Lambda_{w_0}]_{m,l}}{\|H_i \Lambda_{w_0} x_0\|^2 + c_2/L} \tag{3.4}$$

$$= \left[\Lambda_{w_0} \left(Nc_1 \sum_{i=0}^{P-1} \frac{H_i^{\mathrm{T}} H_i}{\|H_i \Lambda_{w_0} x_0\|^2 + c_2/L} \right) \Lambda_{w_0} \right]_{m,l}, \tag{3.5}$$

$$m,l = 0, 1, \ldots, N-1. \tag{3.6}$$

It is worth noticing that the sensitivity matrix M is a cropped version (first N lines and N columns) of a larger, L-by-L matrix, which in turn is a sum of circulant matrices multiplied by Λ_{w_0} on the left and on the right. That multiplication on both sides produces a windowing (a point-wise multiplication) of the inner matrix (sum of circulant matrices) by the separable window $w_0(m,l) = w_0(m)\, w_0(l),\ m,l = 0, 1, \ldots, L-1$.

3.1.1 Case $L = N$

We will now treat the case where the DFT size matches the signal size, $L = N$. As we shall see, the case $L \neq N$ reduces to this first one, if N is sufficiently large. With this condition, the sensitivity matrix becomes (note that now the diagonal and the circulant matrices are N-by-N)

$$M(x) = \Lambda_w \left(Nc_1 \sum_{i=0}^{P-1} \frac{H_i^{\mathrm{T}} H_i}{\|H_i \Lambda_w x\|^2 + c_2/N} \right) \Lambda_w, \tag{3.7}$$

which, when defining the unwindowed sensitivity matrix $M_c(x)$ by

$$M_c(x) = Nc_1 \sum_{i=0}^{P-1} \frac{H_i^{\mathrm{T}} H_i}{\|H_i \Lambda_w x\|^2 + c_2/N}, \tag{3.8}$$

[1] The order of differentiation does not matter as $d(x,y) \in C^2(\mathbb{R}^N)$ with respect to y.

produces

$$M(x) = \Lambda_w M_c(x) \Lambda_w. \tag{3.9}$$

As H_i is a circulant matrix, $H_i^{\mathrm{T}} H_i$ is circulant as well (it is the auto-correlation matrix of h_i) and $M_c(x)$ ends up belonging also to that class (note that the sum or product of circulant matrices is circulant as well).

It is known from the theory of circulant matrices that these are diagonalized by the DFT matrix [25], given by[2]

$$[D_N]_{m,l} = \frac{1}{\sqrt{N}} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}ml}, \quad m,l = 0, 1, \ldots, N-1. \tag{3.10}$$

A circulant matrix V is then similar to a diagonal matrix $\Lambda_{\hat{v}}$ and is expressed as ($^{\mathrm{H}}$ denotes conjugate transposition)

$$V = D_N^{\mathrm{H}} \Lambda_{\hat{v}} D_N, \tag{3.11}$$

where the diagonal matrix is obtained by taking the unnormalized DFT of the first column of V, as in Equation (2.43) (with $L = N$). This knowledge enables us to diagonalize H_i and $H_i^{\mathrm{T}} H_i$ as

$$H_i = D_N^{\mathrm{H}} \Lambda_{\hat{h}_i} D_N \quad \text{and} \tag{3.12}$$

$$H_i^{\mathrm{T}} H_i = H_i^{\mathrm{H}} H_i = D_N^{\mathrm{H}} \Lambda_{\hat{h}_i^*} D_N D_N^{\mathrm{H}} \Lambda_{\hat{h}_i} D_N = D_N^{\mathrm{H}} \Lambda_{|\hat{h}_i|^2} D_N, \tag{3.13}$$

respectively, where we used the fact that D_N is a unitary matrix (defining I as the identity matrix, we have $D_N D_N^{\mathrm{H}} = D_N^{\mathrm{H}} D_N = \mathrm{I}$). Note that we denoted the component-wise absolute value of a vector v by $|v|$, the component-wise conjugation by v^* and the component-wise exponentiation to a power $n \in \mathbb{Z}$ by v^n. We can thus finally diagonalize M_c with simple algebraic manipulations, namely

$$\begin{aligned} M_c(x) &= N c_1 \sum_{i=0}^{P-1} \frac{D_N^{\mathrm{H}} \Lambda_{|\hat{h}_i|^2} D_N}{\|H_i \Lambda_w x\|^2 + c_2/N} && (3.14) \\ &= D_N^{\mathrm{H}} \left(N c_1 \sum_{i=0}^{P-1} \frac{\Lambda_{|\hat{h}_i|^2}}{\|H_i \Lambda_w x\|^2 + c_2/N} \right) D_N && (3.15) \\ &= D_N^{\mathrm{H}} \left(N c_1 \sum_{i=0}^{P-1} \frac{\Lambda_{|\hat{h}_i|^2}}{\frac{1}{N} \sum_{f=0}^{N-1} |\hat{h}_i(f)|^2 \, |\widehat{xw}(f)|^2 + \frac{1}{N} c_2} \right) D_N && (3.16) \\ &= D_N^{\mathrm{H}} \Lambda_{N\hat{a}(x)^2} D_N, && (3.17) \end{aligned}$$

meaning that the eigenvalues of $M_c(x)$ are simply found to be, up to a scaling factor of N, the inverse of the masking threshold $\hat{a}^2$ at the frequency grid $f/N \cdot f_s,\ f = 0, 1, \ldots, N-1$. Furthermore, being the unnormalized DFT of the first column of $M_c(x)$ at point f given by $N\hat{a}^2(x, f)$, the first

[2] For convenience, the normalized DFT matrix is used here.

column itself is given by the inverse DFT

$$[M_c(x)]_{m,0} = \frac{1}{N}\sum_{f=0}^{N-1} N\hat{a}^2(x,f)\,\mathrm{e}^{\mathrm{j}\frac{2\pi}{N}fm} = \sum_{f=0}^{N-1}\hat{a}^2(x,f)\,\mathrm{e}^{\mathrm{j}\frac{2\pi}{N}fm}, \quad m,l = 0,1,\ldots,N-1 \qquad (3.18)$$

and, for any column, the unwidowed sensitivity matrix results in

$$[M_c(x)]_{m,l} = [M_c(x)]_{(m-l \bmod N),0} = \sum_{f=0}^{N-1}\hat{a}^2(x,f)\,\mathrm{e}^{\mathrm{j}\frac{2\pi}{N}f(m-l)}, \quad m,l = 0,1,\ldots,N-1. \qquad (3.19)$$

The sensitivity matrix (3.9), (3.17) can be best interpreted in terms of a variable transformation in the distortion measure (2.49). If we make the substitution

$$z(x) = \sqrt{N}D_N\Lambda_w x, \qquad (3.20)$$

then, as proved in [48], the distortion measure (2.20), (2.42) can be expressed as a quadratic form in z,

$$d(x,y) \approx (z(y)-z(x))^{\mathrm{H}}M_z(z(x))(z(y)-z(x)), \qquad (3.21)$$

with a sensitivity matrix

$$M_z(z) = (\sqrt{N}D_N\Lambda_w)^{-\mathrm{H}}M(x(z))(\sqrt{N}D_N\Lambda_w)^{-1} = \Lambda_{\hat{a}(x(z))^2}, \qquad (3.22)$$

where[3]

$$x(z) = \Lambda_w^{-1}\frac{D_N^{\mathrm{H}}}{\sqrt{N}}z \qquad (3.23)$$

is equal to the inverse function of $z(x)$ of Equation (3.20). This variable transformation has the intuitive meaning that, in the windowed frequency domain $z(x)$, the sensitivity of the distortion measure to each frequency bin is the inverse of the masking threshold $\hat{a}(x(z))^2$ at that bin, i.e., the weighting of Equation (2.42). Furthermore, there is no mutual interaction between different frequency components, as the sum (2.42) treats each term independently.

3.1.2 Case $L \neq N$, L/N integer

In the case $L \neq N$ with L/N integer[4], the sensitivity matrix is not circulant up to a windowing, as in the case $L = N$. Nevertheless, we can define $M_e(x)$ as the L-by-L uncropped version of the it,

$$M_e(x) = \Lambda_{w_0}\left(Nc_1\sum_{i=0}^{P-1}\frac{H_i^{\mathrm{T}}H_i}{\|H_i\Lambda_{w_0}x_0\|^2 + c_2/L}\right)\Lambda_{w_0}, \qquad (3.24)$$

[3] We admit $w(n) \neq 0$, $\forall n$ so that Λ_w is invertible.

[4] We will not treat the non-integer case due to its lack of interest: as N and L are usually powers of two due to the increased computational efficiency when performing Fast Fourier Transforms (FFTs), L/N is usually integer.

and again $M_c(x)$ as the unwindowed version of this last matrix,

$$M_c(x) = Nc_1 \sum_{i=0}^{P-1} \frac{H_i^{\mathrm{T}} H_i}{\|H_i \Lambda_{w_0} x_0\|^2 + c_2/L}, \tag{3.25}$$

which is circulant. To calculate the N-by-N snippet of interest of $M_e(x)$ explicitly, we tile $M_c(x)$ in four sub-matrices, making the size of the first one N-by-N. More formally, we use the representation

$$M_c(x) = \begin{bmatrix} M_t & M_{t2} \\ M_{t3} & M_{t4} \end{bmatrix}, \tag{3.26}$$

where the dimensions of $M_{\{t,t2,t3,t4\}}$ are N-by-N, N-by-$(L-N)$, $(L-N)$-by-N and $(L-N)$-by-$(L-N)$, respectively. Notice that the sub-matrices are Toeplitz (but not circulant). We can now proceed to the calculation of M_e in terms of M_t by performing multiplication of block matrices. If we consider $0_{m,l}$ to be an m-by-l zero matrix, we get

$$\begin{aligned} M_e(x) &= \begin{bmatrix} \Lambda_w & 0_{N,L-N} \\ 0_{L-N,N} & 0_{L-N,L-N} \end{bmatrix} \begin{bmatrix} M_t & M_{t2} \\ M_{t3} & M_{t4} \end{bmatrix} \begin{bmatrix} \Lambda_w & 0_{N,L-N} \\ 0_{L-N,N} & 0_{L-N,L-N} \end{bmatrix} && (3.27) \\ &= \begin{bmatrix} \Lambda_w M_t & \Lambda_w M_{t2} \\ 0_{L-N,N} & 0_{L-N,L-N} \end{bmatrix} \begin{bmatrix} \Lambda_w & 0_{N,L-N} \\ 0_{L-N,N} & 0_{L-N,L-N} \end{bmatrix} && (3.28) \\ &= \begin{bmatrix} \Lambda_w M_t \Lambda_w & 0_{N,L-N} \\ 0_{L-N,N} & 0_{L-N,L-N} \end{bmatrix}. && (3.29) \end{aligned}$$

From these calculations, it follows that $M(x)$ can be seen as a Toeplitz matrix $M_t(x)$ which has been windowed by the separable window $w(m,l) = w(m)\,w(l)$. That Toeplitz matrix is itself an N-by-N snippet of the larger circulant matrix $M_c(x)$. Mathematically

$$M(x) = \Lambda_w M_t(x) \Lambda_w, \tag{3.30}$$

where

$$[M_t(x)]_{m,l} = [M_c(x)]_{m,l}, \quad m,l = 0,1,\ldots,N-1. \tag{3.31}$$

Calculating the determinant, inverse and other functions of a Toeplitz matrix is, in general, a mathematically intractable task. Nevertheless, there are results on the asymptotic ($N \to \infty$) behavior of this class of matrices, which approximate a Toeplitz matrix by a carefully chosen circulant one with the same asymptotic characteristics. Indeed, it was shown in [25] that, if we admit that the sequence built from M_t with $N \to \infty$ given by

$$t(m) = \begin{cases} [M_t(x)]_{0,-m}, & m = \ldots,-3,-2,-1 \\ [M_t(x)]_{m,0}, & m = 0,1,2,\ldots \end{cases} \tag{3.32}$$

is absolutely summable, then we can build a circulant matrix $\bar{M}_c(x)$,

$$\bar{M}_c(x) = D_N^{\mathrm{H}} \Lambda_{e_c} D_N, \tag{3.33}$$

with eigenvalues given by

$$e_c(f) = \sum_{m=-N+1}^{N-1} t(m)\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm}, \quad f = 0, 1, \ldots, N-1 \tag{3.34}$$

that is asymptotically equivalent to $M_t(x)$ in the sense that the *Hilbert-Schmidt norm* (also called the *weak norm*) of the difference

$$\|\bar{M}_c(x) - M_t(x)\|_{\mathrm{HS}} = \sqrt{\frac{1}{N} \sum_{m=0}^{N-1} \sum_{l=0}^{N-1} \left([\bar{M}_c(x)]_{m,l} - [M_t(x)]_{m,l}\right)^2} \tag{3.35}$$

vanishes with increasing N. To make the approximation $\bar{M}_c(x)$ compatible with the calculations for $L = N$ (we want to transform the case $L \neq N$ in a slightly modified version of the case $L = N$), we define an "equivalent" inverse of the masking threshold $\hat{\bar{a}}^2(x, f)$ in terms of the eigenvalues $e_c(\cdot)$ as (cf. Equations (3.17), (3.33))

$$N\hat{\bar{a}}^2(x, f) \overset{\mathrm{def}}{=} e_c(f). \tag{3.36}$$

We can thus express the approximation for the unwindowed sensitivity matrix as

$$\bar{M}_c(x) = D_N^{\mathrm{H}} \Lambda_{N\bar{a}^2(x)} D_N \tag{3.37}$$

and, following the same steps of (3.18), (3.19) and of the corresponding discussion, also as

$$[\bar{M}_c(x)]_{m,l} = \sum_{f=0}^{N-1} \hat{\bar{a}}^2(x, f)\, \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}f(m-l)}, \quad m,l = 0, 1, \ldots, N-1. \tag{3.38}$$

We can particularize the expression for the eigenvalues of $\bar{M}_c(x)$ by mapping the values of $t(m)$ in (3.32) to the corresponding ones in $[M_c(x)]_{\cdot,0}$:

$$\begin{aligned}
e_c(f) &= \sum_{m=-N+1}^{-1} [M_c(x)]_{m+L,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} + \sum_{m=0}^{N-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} && (3.39)\\
&= \sum_{m=L-N+1}^{L-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}f(m-L)} + \sum_{m=0}^{N-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} && (3.40)\\
&= \mathrm{e}^{\mathrm{j}2\pi\frac{L}{N}f} \left(\sum_{m=L-N+1}^{L-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} \right) + \sum_{m=0}^{N-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} && (3.41)\\
&= \sum_{m=L-N+1}^{L-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} + \sum_{m=0}^{N-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} && (3.42)
\end{aligned}$$

$$e_c(f) = \sum_{m=0}^{L-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} - \sum_{m=N}^{L-N} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm}, \tag{3.43}$$

where in (3.42) we used the fact that L/N is integer. To interpret the last equation better, it is useful to calculate what the values of $[M_c(x)]_{\cdot,0}$ are. Following the same procedure as the one in equations (3.12) to (3.17) of the case $L = N$, we now have for $L \neq N$ the unwindowed extension of the sensitivity matrix given by

$$\begin{aligned} M_c(x) &= D_L^{\mathrm{H}} \left(N c_1 \sum_{i=0}^{P-1} \frac{\Lambda_{|\hat{h}_i|^2}}{\|H_i \Lambda_{w_0} x_0\|^2 + c_2/L} \right) D_L & (3.44) \\ &= D_L^{\mathrm{H}} \Lambda_{L\hat{a}(x)^2} D_L, & (3.45) \end{aligned}$$

and its first column by

$$[M_c(x)]_{m,0} = \frac{1}{L} \sum_{f=0}^{L-1} L\hat{a}^2(x,f)\, \mathrm{e}^{\mathrm{j}\frac{2\pi}{L}fm} = \sum_{f=0}^{L-1} \hat{a}^2(x,f)\, \mathrm{e}^{\mathrm{j}\frac{2\pi}{L}fm}, \quad f = 0, 1, \dots, L-1. \tag{3.46}$$

Equation (3.46) tells us that $[M_c(x)]_{\cdot,0}$ are the (circulant) autocorrelation samples of the inverse of the masking threshold in the time domain $a(x, \cdot)$, up to a scaling factor. For sufficiently large N, (empirically values starting from $N \sim 1000$ at $f_s = 48$ kHz are acceptable), the masking threshold will be practically uncorrelated with itself at lags greater than N (and smaller than $L - N$, as the problem is defined modulo L), being as consequence the second term in (3.43) negligible with respect to the first one. The "equivalent" inverse of the masking threshold associated with the Toeplitz matrix $M_t(x)$ is then

$$\begin{aligned} \hat{\hat{a}}^2(x,f) &\approx \frac{1}{N} \sum_{m=0}^{L-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}fm} & (3.47) \\ &= \frac{L}{N} \left(\frac{1}{L} \sum_{m=0}^{L-1} [M_c(x)]_{m,0}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{L}\left(\frac{L}{N}f\right)m} \right) & (3.48) \\ &= \frac{L}{N}\, \hat{a}^2\left(x, \frac{L}{N}f\right), \quad f = 0, 1, \dots, N-1, & (3.49) \end{aligned}$$

where in (3.49) we performed the DFT of (3.46) and used again the fact that L/N is integer. As a final step we can also approximate the squared magnitudes of the band-pass filters $\hat{h}_i$ in $\hat{a}^2(x, Lf/N)$ by its decimated versions when their impulse response is essentially limited to N samples (as above, N starting from around 1000 samples at 48 kHz gives a good approximation). The equivalent inverse of the masking threshold becomes

$$\hat{\hat{a}}^2(x,f) = \frac{L}{N}\hat{a}^2\left(x, \frac{L}{N}f\right) = N c_1 \sum_{i=0}^{P-1} \frac{\frac{L}{N}|\hat{h}_i(\frac{L}{N}f)|^2}{\sum_{f'=0}^{L-1} |\hat{h}_i(f')|^2\, |\widehat{xw}(f')|^2 + c_2} \tag{3.50}$$

$$\hat{\bar{a}}^2(x,f) \approx Nc_1 \sum_{i=0}^{P-1} \frac{\frac{L}{N}|\hat{h}_i(\frac{L}{N}f)|^2}{\frac{L}{N}\sum_{f'=0}^{N-1}|\hat{h}_i(\frac{L}{N}f')|^2\,|\widehat{xw}_N(f')|^2+c_2} \tag{3.51}$$

$$= Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(f)|^2}{\sum_{f'=0}^{N-1}|\hat{\bar{h}}_i(f')|^2\,|\widehat{xw}_N(f')|^2+c_2},$$

$$f = 0,1,\ldots,N-1, \tag{3.52}$$

where the subscript N denotes the execution of an N-point DFT instead of an L-point one (Equation (2.43) with $L=N$) and

$$\hat{\bar{h}}_i(f) \overset{\text{def}}{=} \sqrt{\frac{L}{N}}\,\hat{h}_i\left(\frac{L}{N}f\right), \quad f = 0,1,\ldots,N-1. \tag{3.53}$$

In the remaining of this paper, we will always use the circulant approximation $\bar{M}_c(x)$, the equivalent inverse of the masking threshold $\hat{\bar{a}}^2$ and the eigenvalues $N\hat{\bar{a}}^2(x,\cdot)$ of the equivalent matrix when we deal with the case $L \neq N$, i.e., we will be using

$$\bar{M}(x) = \Lambda_w \bar{M}_c(x)\Lambda_w \tag{3.54}$$

and equations (3.37), (3.52) and (3.53) instead of the exact $M(x)$ for $L \neq N$ of Equation (3.30), so that this case degenerates in the case $L = N$ by using $\bar{M}_c(x)$ as the circulant matrix in (3.9). Although the theory related to this approximation does not state anything on the individual values of the matrices $\bar{M}_c(x)$ and $M_t(x)$ or on individual eigenvalues [25], it guarantees two asymptotic matches. In the first place, as discussed previously, the "distance" between the matrices, measured in the Hilbert-Schmidt norm, disappears asymptotically. In the second place, if we define the Fourier Series[5] as the limit of (3.34) when $N \to \infty$,

$$t_{\text{FS}}(\lambda) = \sum_{m=-\infty}^{\infty} t(m)\,\mathrm{e}^{-\mathrm{j}m\lambda}, \tag{3.55}$$

and an arbitrary continuous real function Ψ on $[\inf_\lambda t_{\text{FS}}(\lambda); \sup_\lambda t_{\text{FS}}(\lambda)]$, we know that the eigenvalues $e_t(f), f = 0,1,\ldots,N-1$ of $M_t(x)$ are bounded by the limits of the mentioned interval (being the bound asymptotically tight), and that the average of the function Ψ of the eigenvalues is asymptotically the same on both matrices $M_t(x)$ and $\bar{M}_c(x)$, i.e.,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{f=0}^{N-1}\Psi(e_t(f)) = \lim_{N\to\infty}\frac{1}{N}\sum_{f=0}^{N-1}\Psi(e_c(f)). \tag{3.56}$$

We say in this case that the eigenvalues $e_t(\cdot)$ and $e_c(\cdot)$ are *asymptotically equally distributed.*

[5] Remember that we admitted that $t(m)$ is absolutely summable, so that t_{FS} exists.

3.2 A Suboptimal Compressor

Having discovered the form of the sensitivity matrix in Section 3.1, it is now natural to work out more deeply what condition the optimal compressor should satisfy, check whether optimality can be achieved or not and, finally, develop a compressor using the derived condition. In this section we do precisely that work, showing that no optimal companding scheme exists upon the imposition of certain conditions and deriving an expression for a suboptimal compressor using some of the equations regarding the optimality condition.

As explained intuitively in Section 2.2 and extensively in [38], a compander $\{F, F^{-1}\}$ is optimal if and only if its Jacobian matrix $F'(x)$ satisfies (2.19) up to a positive real scaling factor, where the equation has to be satisfied with probability one (i.e., it can be left unsatisfied for certain values of x if they never occur) and $M(x)$ is the sensitivity matrix of Equation (2.5). Without loss of generality, we set that scaling factor to one here. In addition, as explained extensively in Section 3.1, the sensitivity matrix is given, as an approximation, by

$$M(x) = \Lambda_w D_N^{\mathrm{H}} \Lambda_{N\hat{\hat{a}}^2(x)} D_N \Lambda_w, \tag{3.57}$$

where $\hat{\hat{a}}^2(x)$, Equation (3.52), is an N-point approximation for the decimated L-point inverse of the masking threshold. When $L = N$ these results are exact.

As stated in Subsection 2.2.4, we call any matrix $F'(x)$ satisfying (2.19) a *square-root* of the matrix $M(x)$, and denote a particular square-root by $\sqrt{M(x)}$. Due to $M(x)$ degenerating in a very simple windowed circulant matrix, which in turn is diagonalized by the DFT matrix, one possible square-root is very easy to calculate. Indeed try this square-root:

$$F'(x) = \sqrt{M(x)} = D_N^{\mathrm{H}} \Lambda_{\sqrt{N}\hat{\hat{a}}(x)} D_N \Lambda_w. \tag{3.58}$$

To check this solution, we must prove in the first place that this $\sqrt{M(x)}$ is real, because F is a real-valued function of real argument. This happens when the matrix conjugate of $\sqrt{M(x)}$ equals to the matrix itself. Knowing that a triple (normalized) DFT is equivalent to an inverse DFT and vice-versa [45], that the DFT matrix is symmetric, using simple properties of matrix conjugation (which we denote by *) and using the fact that w and $\hat{\hat{a}}$ are real, we get

$$\left(\sqrt{M(x)}\right)^* = D_N^{\mathrm{T}} \Lambda^*_{\sqrt{N}\hat{\hat{a}}(x)} D_N^* \Lambda_w^* = D_N^{\mathrm{H}} \left({D_N^2}^{\mathrm{H}} \Lambda_{\sqrt{N}\hat{\hat{a}}(x)} D_N^2 \right) D_N \Lambda_w. \tag{3.59}$$

It is easy to see through the duality property of the DFT (e.g., [45]) that

$$D_N^2 = {D_N^2}^{\mathrm{H}} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & \vdots & ⋰ & \vdots & \vdots \\ 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{3.60}$$

is the reverse operator, i.e., $[D_N^2 v]_n = v((N-n) \bmod N),\ n = 0, 1, \dots, N-1$ for any vector $v = [v_0, v_1, \dots, v_{N-1}]^{\mathrm{T}}$. Furthermore, if you left and right multiply a diagonal matrix Λ_v by two reverse operators, you get again a diagonal matrix, but with the reversed vector:

$$D_N^2 \Lambda_v D_N^2 = \Lambda_{D_N^2 v}. \tag{3.61}$$

Applying Equation (3.61) to (3.59) produces

$$\left(\sqrt{M(x)}\right)^* = D_N^{\mathrm{H}} \Lambda_{\sqrt{N} D_N^2 \hat{a}(x)} D_N \Lambda_w = D_N^{\mathrm{H}} \Lambda_{\sqrt{N} \hat{a}(x)} D_N \Lambda_w = \sqrt{M(x)}, \tag{3.62}$$

where we took into account that, due to the dependence of $\hat{a}$ solely on magnitudes of DFT's of real signals or decimated versions of them (cf. Equations (3.52) and (3.53)), $\hat{a}(x)$ has even symmetry. As wanted, we have thus proven that $\sqrt{M(x)}$ is real.

We will now finally check that this choice of $\sqrt{M(x)}$ fulfills Equation (3.58). Using the fact that $F'(x) = \sqrt{M(x)}$ is real, we get

$$F'(x)^{\mathrm{T}} F'(x) = F'(x)^{\mathrm{H}} F'(x) = \Lambda_w^{\mathrm{H}} D_N^{\mathrm{H}} \Lambda_{\sqrt{N}\hat{a}(x)}^{\mathrm{H}} D_N D_N^{\mathrm{H}} \Lambda_{\sqrt{N}\hat{a}(x)} D_N \Lambda_w = M(x), \tag{3.63}$$

where we used the unitary property of D_N and the fact that w and $\hat{a}$ are real. As was previously noted in Subsection (2.2.4), this solution is not unique. All matrices of the form $U(x)F'(x)$ with an orthogonal $U(x)$ are solution of Equation (2.19), and all solutions are separated by the left multiplication by an orthogonal matrix $U(x)$. Although this orthogonal freedom of $F'(x)$ exists, in this thesis we will only deal with the case $U(x) = \mathrm{I}$. Note anyway that, for U independent of x, a sub-optimal companding scheme exists for the square-root of the sensitivity matrix $\sqrt{M(x)}$ of Equation (3.58) if and only if it exists for $U\sqrt{M(x)}$. The existence of an optimal scheme is thus covered for all cases where U is independent of x, even setting $U = \mathrm{I}$.

Also for U independent of x, note that when no optimal companding scheme exists and when we build a sub-optimal compander $\{\tilde{F}, \tilde{F}^{-1}\}$ for the square-root of the sensitivity matrix $\sqrt{M(x)}$ of Equation (3.58), we can also build a sub-optimal compander $\{\tilde{F}_2, \tilde{F}_2^{-1}\}$ for the square-root $U\sqrt{M(x)}$ by doing $\tilde{F}_2(x) = U\tilde{F}(x)$ and $\tilde{F}_2^{-1}(\xi) = \tilde{F}^{-1}(U^{-1}\xi)$. Nevertheless, we can see through equations (2.38) and (2.39) that no performance is gained (or lost) by doing this. More clever solutions have to be employed to get a performance difference for the square-root $U\sqrt{M(x)}$.

We now have a condition for the optimality of the scheme in terms of the Jacobian matrix of F (Equation (3.58)). Motivated by the variable substitution formulation of [48] and by the specific one of Section 3.1, we introduce a new function G, written in terms of F as

$$F(x) = \frac{D_N^H}{\sqrt{N}}\, G(\sqrt{N} D_N \Lambda_w x). \tag{3.64}$$

This expression can be best seen as block diagram, as depicted in Figure 3.1. First we point-wise multiply the input signal x by the window w (× denotes point-wise multiplication), we take an unnormalized DFT, we then apply a function G (to be developed), and finally emit the inverse DFT as our compressed signal, $F(x)$. In other words, we process the windowed signal in the frequency domain instead of the original signal directly in the time domain, using thus the variable substitution of Equation (3.20).

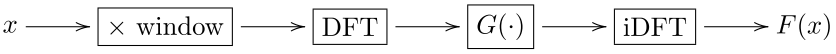


Figure 3.1: Block Diagram of the compressor $F(x)$.

As G operates on complex vectors and emits complex vectors (DFTs of real signals), it is a complex-valued complex function. If we assume it to be complex differentiable on all its variables in an open set, we can use the chain rule [5] to get the Jacobian matrix of F in terms of the one of G:

$$F'(x) = D_N^H\, G'(\sqrt{N} D_N \Lambda_w x)\, D_N \Lambda_w. \tag{3.65}$$

By simple comparison with (3.58), and using the variable substitution (3.20), we get the optimality condition in terms of G' as

$$G'(z) = \Lambda_{\sqrt{N}\hat{\hat{a}}(x(z))} \equiv \Lambda_{\sqrt{N}\hat{\hat{a}}(z)}, \tag{3.66}$$

where $x(z)$ is the inverse function of $z(x)$, Equation (3.23), and we do a slight abuse of notation, saying that now $\hat{\hat{a}}$ is a function of z directly.

If we write down Equation (3.66) explicitly, making use of (3.52), we get as a result this equation in branches:

$$\frac{\partial G_m}{\partial z_l}(z) = \begin{cases} \sqrt{N}\sqrt{Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\hat{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\hat{h}}_i(f)|^2\,|z(f)|^2 + c_2}} & \text{for } l = m \\ 0 & \text{for } l \neq m \end{cases}, \quad m,l = 0,1,\ldots,N-1. \tag{3.67}$$

Equation (3.67) makes a severe constraint visible, which appears when we want to satisfy the optimality conditions with the assumptions that we have made. Remember that we assumed G

to be complex differentiable (on all variables) and that, from complex analysis, we know that any complex differentiable function defined on an open set has all its derivatives and antiderivatives (e.g., [13]), due to being an analytic function (this relation is actually an equivalence). A consequence of this sentence, of our assumption and of condition (3.67) is that all elements of $\sqrt{M}$ (right hand side of Equation (3.67)) have to be differentiable with respect to all variables. It is easy to see, due to the dependence of $\hat{a}(z)$ on $|z(n)|^2$, $\forall n$ (use the Cauchy-Riemann equations), that this function is not differentiable in any open set on any of its coordinates, implying that no differentiable G (in an open set) exists which satisfies Equation (3.67) directly.

There are two ways out of this impasse: either we do not assume G to be analytic and use the theory of the complex differential forms $\partial/(\partial z)$ and $\partial/(\partial z^*)$ [8], associated with possibly non-analytic complex functions which are differentiable when seen as real functions with the double number of arguments, or we perform a very simple substitution. As z, the argument of G, is the DFT of a real signal, the hermitian symmetry property applies [45]:

$$z^*(f) = z(N-f), \quad f = 0, 1, \ldots, N-1; \quad z(N) \equiv z(0). \tag{3.68}$$

If we substitute $z^*(f)$ in Equation (3.67) (remember that $|z(f)|^2 = z(f)z^*(f)$), we end up with analytic functions on the right-hand side, namely

$$\frac{\partial G_m}{\partial z_l}(z) = \begin{cases} \sqrt{N}\sqrt{Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\hat{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\hat{h}}_i(f)|^2 \, z(f)z(N-f)+c_2}} & \text{for } l = m \\ 0 & \text{for } l \neq m \end{cases}, \quad m,l = 0, 1, \ldots, N-1. \tag{3.69}$$

These functions are analytic due to being compositions, sums, quotients and products of functions which are analytic themselves. It should be noted that, in the domain of interest, corresponding to Equation (3.68), the argument $\cdot$ of the functions $1/\cdot$ and $\sqrt{\cdot}$ in Equation (3.69) is real and strictly positive (the strictness comes from the positive real constant c_2), so that indeed we are on the domain of analyticity of those functions. We use the conventional definition of the argument of a complex number in the interval $]-\pi,\pi]$, so that the domain of analyticity is $\mathbb{C}\backslash\{x : x \in \mathbb{R}_0^-\}$.

It is possible to derive a function G making use of the mentioned theory of the complex differential forms and, if we would make such a derivation, due to restrictions made on that function on the kind of vectors it gets and it receives (vectors with hermitian symmetry, due to the fact that their time-domain counterpart is real-valued), we would arrive at the same final result as with this simple substitution. Due to the lack of space in this thesis paper, this derivation will not be in the scope of the it, although it was made during the working period of the thesis. We will thus make all derivations in this section taking the substitution into account.

As expected (see Subsection 2.2.4), the problem of equations (3.66), (3.69) is over-determined and to have a solution (assuming now G analytic and using the substitution related to the hermitian symmetry of z) it is required that

$$\frac{\partial[\Lambda_{\sqrt{N}\hat{a}(z)}]_{m,l}}{\partial z_k} = \frac{\partial[\Lambda_{\sqrt{N}\hat{a}(z)}]_{m,k}}{\partial z_l}, \quad m,l,k = 0,1,\ldots,N-1. \tag{3.70}$$

It can be seen easily that, as $\Lambda_{\sqrt{N}\hat{a}}$ is a diagonal matrix, this constriction is not satisfied when $l = m$ and $k \neq m$. Indeed, $[\Lambda_{\sqrt{N}\hat{a}}]_{m,k}$ is 0, so that its derivative with respect to z_m is also 0, but $[\Lambda_{\sqrt{N}\hat{a}}]_{m,m}$ depends on all variables, being its derivative with respect to z_k not 0. Even if we do not assume G to be analytic (but at least of class $C^1(\mathbb{R}^{2N})$, i.e., continuously differentiable when seen as real function), it is possible to prove, using the theory of the complex differential forms, that this problem does not have a solution when making natural restrictions on G relative to the hermitian symmetry of its input and output. As a consequence, no optimal compressor exists for finite N for any signal independent orthogonal matrix U in (2.33). The case of signal dependent orthogonal matrices $U(x)$ is left open for future work.

We will thus proceed building a suboptimal compressor (and its correspondent expander), which only satisfies some of the equations (3.69). Obviously we do not want a constant, signal independent compressor, so that we choose to satisfy the equations on the diagonal of the matrices in Equation (3.66). As the functions on the right-hand side are analytic (cf. Equation (3.69)), we can employ the Fundamental Theorem of Calculus [13] to solve them. Furthermore, we admit that N is even, since the most interesting values of N for the execution of the Fast Fourier Transform (FFT), a dominantly used algorithm for the execution of the DFT, are powers of two. Using this assumption, we define the suboptimal compressor as

$$\tilde{G}_m(z) = \sqrt{N} \int_0^{z(m)} \sqrt{N c_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{\substack{f=0\\ f\neq m,N-m}}^{N-1} |\hat{\bar{h}}_i(f)|^2\, z(f) z(N-f) + 2|\hat{\bar{h}}_i(m)|^2\, z(N-m) v + c_2}}\, \mathrm{d}v \tag{3.71}$$

$$= \sqrt{N} \int_0^{z(m)} \sqrt{N c_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\, |z(f)|^2 + 2|\hat{\bar{h}}_i(m)|^2\, z^*(m)\, [v - z(m)] + c_2}}\, \mathrm{d}v \tag{3.72}$$

for $m \neq 0$ and $m \neq N/2$, and as

$$\tilde{G}_m(z) = \sqrt{N} \int_0^{z(m)} \sqrt{N c_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{\substack{f=0\\ f\neq m}}^{N-1} |\hat{\bar{h}}_i(f)|^2\, z(f) z(N-f) + |\hat{\bar{h}}_i(m)|^2\, v^2 + c_2}}\, \mathrm{d}v \tag{3.73}$$

$$= \sqrt{N} \int_0^{z(m)} \sqrt{N c_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\, |z(f)|^2 + |\hat{\bar{h}}_i(m)|^2\, [v^2 - z^2(m)] + c_2}}\, \mathrm{d}v \tag{3.74}$$

for $m = 0$ or $m = N/2$ (notice that $z(m)$ is real in this case), where the integral is a complex path integral and the integration path is arbitrary, provided that it remains within the domain of analyticity of $\hat{\bar{a}}$. For fulfilling the upper equations of (3.69), the lower limit of the integral can also be arbitrary, but here we choose it to be 0. This choice has the following explanation. As the input signal z is a DFT vector, it is reasonable to assume that its components are on average 0 (except eventually for the DC component, but that component is perceptually irrelevant so that it can be filtered out before the whole source-coding chain). When calculating the Jacobian matrix of this suboptimal compressor, the out-diagonal elements of it are also given in terms of a complex integral with the same limits (with a partial derivative as the integrand). As we would want those components ideally to be 0 (lower equations of (3.69)), a natural approach is to make the actual components to be, on average, as close as 0 (in terms of their magnitude) as possible. Choosing then the lower limit of the integral to be equal to the expected value of its upper limit (0, in this case), we minimize the length run by the path integral (on average), so that the magnitude of the integrals corresponding to the out-diagonal derivatives is also, on average, minimized.

For the practical implementation, the path choice $v = t\, z(m),\ t \in [0;1]$ is convenient (as before, the argument of $1/\cdot$ and $\sqrt{\cdot}$ never becomes real non-positive for any t, so that this path choice is valid), because this choice leads to the calculation of a real integral of a real function:

$$\tilde{G}_m(z) = \sqrt{N} z(m) \int_0^1 \sqrt{N c_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\, |z(f)|^2 + 2|\hat{\bar{h}}_i(m)|^2\, |z(m)|^2\, (t-1) + c_2}}\, \mathrm{d}t \tag{3.75}$$

for $m \neq 0$ and $m \neq N/2$, and

$$\tilde{G}_m(z) = \sqrt{N} z(m) \int_0^1 \sqrt{N c_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\, |z(f)|^2 + |\hat{\bar{h}}_i(m)|^2 z^2(m)\, (t^2-1) + c_2}}\, \mathrm{d}t \tag{3.76}$$

for $m = 0$ or $m = N/2$. The expression (3.75)/(3.76) is also convenient to analyze a key requirement of the compressor: its output must be hermitian symmetric as we want a real compressed signal in the time-domain. In this form, we can see that the dependence of the integral on m is only on the magnitude of hermitian symmetric signals. For that reason, the result of the integration is the same for $\tilde{G}_m$ and $\tilde{G}_{N-m}$. Obviously, as the integrand is real and positive, the integral itself is also real and positive, so that we can see the compression operation as the point-wise multiplication of z by a certain real positive gain (the integral), that we will label as $\Gamma_m(z)$, i.e., we can see the compressor as

$$\tilde{G}_m(z) = \sqrt{N}\, \Gamma_m(z)\, z(m), \quad m = 0, 1, \ldots, N-1 \tag{3.77}$$

with

$$\Gamma_m(z) = \begin{cases} \int_0^1 \sqrt{Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2 \, |z(f)|^2 + 2|\hat{\bar{h}}_i(m)|^2 \, |z(m)|^2 (t-1) + c_2}} \, \mathrm{d}t & \text{for } m \neq 0, N/2 \\ \int_0^1 \sqrt{Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2 \, |z(f)|^2 + |\hat{\bar{h}}_i(m)|^2 z^2(m)(t^2-1) + c_2}} \, \mathrm{d}t & \text{for } m = 0, N/2. \end{cases} \tag{3.78}$$

As we are point-wise multiplying two signals with hermitian symmetry (one of them is even real), the result has hermitian symmetry as well, as desired.

3.3 Taylor Expansion of the Suboptimal Compressor

Although the developed compressor of Section 3.2 is well defined by equations (3.77) and (3.78), its practical computation with a numerical method is quite expensive. If, for example, we use the adaptive Simpson's method [33], the integration interval has to be split recursively and adaptively until the error between the integral of the quadratic interpolation of an interval directly and the sum of the integrals of the quadratic interpolations of two sub-intervals is small enough. This process can deploy unnecessary function evaluations, specially when the function is itself very well approximable by a quadratic, linear or a constant polynomial in the whole region of integration. In that case, it is preferable to do the approximation analytically and compute directly the value of the integral with the approximation.

In this section, we motivate the approximation of the integrands of (3.78) by their Taylor expansions up to some order M and do the necessary calculations to obtain an expression for the compressor (3.77) in terms of those expansions. As we will see in the simulations (Section 4.3), the needed value of M is very low (even $M = 0$ delivers a good approximation), so that the application of a numerical method with several function evaluations becomes unnecessary.

Indeed, look at Equation (3.78). The only dependence on the integration variable (use the case $m \neq 0, N/2$ as an example) is on the term $2|\hat{\bar{h}}_i(m)|^2 \, |z(m)|^2(t-1)$, being that term summed side by side with $\|\Lambda_{\hat{\bar{h}}_i} z\|^2$. The factor on the left of $t-1$ corresponds to only two terms of the squared norm of the filtered z, so that when N grows, the dependence on t becomes weaker and weaker, i.e.,

$$2|\hat{\bar{h}}_i(m)|^2 \, |z(m)|^2(t-1) \ll \sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2 \, |z(f)|^2 \quad \forall_{t \in [0;1]}, \text{for } N \gg . \tag{3.79}$$

More formally, the average behavior of the quotient of the two terms is characterized statistically by

$$\mathrm{E}\left\{\frac{2|\hat{\bar{h}}_i(m)|^2\,|Z(m)|^2(t-1)}{\sum_{f=0}^{N-1}|\hat{\bar{h}}_i(f)|^2\,|Z(f)|^2}\right\} = 2(t-1)\mathrm{E}\left\{\frac{|\hat{\bar{h}}_i(m)|^2\,\frac{|Z(m)|^2}{N}}{\frac{1}{N}\sum_{f=0}^{N-1}|\hat{\bar{h}}_i(f)|^2\,|Z(f)|^2}\right\} \tag{3.80}$$

$$= 2(t-1)\frac{1}{N}\mathrm{E}\left\{\frac{|\hat{\bar{h}}_i(m)|^2\,\frac{|Z(m)|^2}{N}}{\frac{1}{N}\|(Xw)\circledast\bar{h}_i\|^2}\right\}, \tag{3.81}$$

where $\circledast$ denotes circulant convolution and where juxtaposition denotes the point-wise multiplication. The argument of the expectation of Equation (3.81) is exactly one component of the (windowed) periodogram of the filtered signal over the total estimated power of the same signal. As in Subsection 2.3.2, we assume that X is wide-sense stationary and ergodic here as well. The denominator of the expected value converges thus to the power of the filtered input signal, $P_{[X*\bar{h}_i]}$, when $N \to \infty$. Furthermore, it is well known (e.g., [45]) that the numerator (the periodogram of the input signal) is an inconsistent estimator (its variance does not depend on N), so that, asymptotically, the variance of the denominator becomes negligible with respect to the variance of the numerator. Consequently, we have

$$\mathrm{E}\left\{\frac{2|\hat{\bar{h}}_i(m)|^2\,|Z(m)|^2(t-1)}{\sum_{f=0}^{N-1}|\hat{\bar{h}}_i(f)|^2\,|Z(f)|^2}\right\} \approx 2(t-1)\frac{\frac{1}{N}\mathrm{E}\left\{|\hat{\bar{h}}_i(m)|^2\,\frac{|Z(m)|^2}{N}\right\}}{P_{[X*\bar{h}_i]}}, \tag{3.82}$$

with $P_{[v]} = \mathrm{E}[V_n^2]$ equal to the power of a signal. It is also known [45] that the periodogram is an asymptotically unbiased estimator for the power spectral density of the input signal X (although not consistent), so that in Equation (3.82), for large N, we are approximately dividing the power of one band (of size $1/N$) of the filtered input signal by the total power of that signal (up to a constant $2(t-1)$). The size of the band decreases with a rate $O(1/N)$ but the power spectrum and signal power remain constant with varying N, so that we achieve

$$\lim_{N\to\infty}\mathrm{E}\left\{\frac{2|\hat{\bar{h}}_i(m)|^2\,|Z(m)|^2(t-1)}{\sum_{f=0}^{N-1}|\hat{\bar{h}}_i(f)|^2\,|Z(f)|^2}\right\} = 0 \tag{3.83}$$

at a rate $O(1/N)$. Furthermore, due to the independence of the variance of the periodogram of Equation (3.82) on N, the variance of the quotient satisfies (note the pre-multiplying factor $1/N$ of Equation (3.81))

$$\lim_{N\to\infty}\mathrm{Var}\left\{\frac{2|\hat{\bar{h}}_i(m)|^2\,|Z(m)|^2(t-1)}{\sum_{f=0}^{N-1}|\hat{\bar{h}}_i(f)|^2\,|Z(f)|^2}\right\} = 0, \tag{3.84}$$

at a rate $O(1/N^2)$. We can thus finally conclude that we have the convergence in probability [16]

$$\lim_{N\to\infty}\frac{2|\hat{\bar{h}}_i(m)|^2\,|Z(m)|^2(t-1)}{\sum_{f=0}^{N-1}|\hat{\bar{h}}_i(f)|^2\,|Z(f)|^2} = 0 \tag{3.85}$$

for all $t \in [0;1]$, which means that the probability that the function of the random variables $Z(m)$ expressed in (3.85) is not 0 can be made arbitrarily small by increasing N. In other words, the probability density function of the left hand side of (3.85) approaches a Dirac delta function at 0 when N goes to infinity.

The limit (3.85) gives us the guarantee that if we build a Taylor expansion of a certain order M for the compressor, the higher the vector size N, the lower the needed order M will be for a good approximation, and that with $N \to \infty$, the order 0 will give the exact result. The construction of the Taylor expansion of the integrand of (3.78) for large N is thus motivated.

Define the integrand of Equation (3.78) as $\gamma_m(z,t)$,

$$\gamma_m(z,t) = \begin{cases} \sqrt{Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\,|z(f)|^2 + 2|\hat{\bar{h}}_i(m)|^2\,|z(m)|^2(t-1)+c_2}} & \text{for } m \neq 0, N/2 \\ \sqrt{Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\,|z(f)|^2 + |\hat{\bar{h}}_i(m)|^2 z^2(m)(t^2-1)+c_2}} & \text{for } m = 0, N/2 \end{cases}, \tag{3.86}$$

noting that

$$\gamma_m(z,1) = \hat{\bar{a}}(z,m), \tag{3.87}$$

and define its Taylor expansion around $t = 1$:

$$\gamma_{m,M}(z,t) = \sum_{k=0}^{M} \frac{1}{k!} \frac{\partial^k \gamma_m}{\partial t^k}(z,1)\,(t-1)^k. \tag{3.88}$$

With this approximation, the compressor gain becomes

$$\Gamma_{m,M}(z) = \int_0^1 \gamma_{m,M}(z,t)\,\mathrm{d}t = \sum_{k=0}^{M} \frac{1}{k!} \frac{\partial^k \gamma_m}{\partial t^k}(z,1) \int_0^1 (t-1)^k\,\mathrm{d}t = \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!} \frac{\partial^k \gamma_m}{\partial t^k}(z,1) \tag{3.89}$$

and the compressor $\tilde{G}_M$ (the subscript M indicates the usage of the Taylor expansion of order M) becomes

$$\tilde{G}_{m,M}(z) = \sqrt{N}\,\Gamma_{m,M}(z)\,z(m) = \sqrt{N} \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!} \frac{\partial^k \gamma_m}{\partial t^k}(z,1)\,z(m). \tag{3.90}$$

Note that with $M = \infty$, this compressor becomes exactly equal to the one developed in Section 3.2, $\tilde{G}(z)$, due to the analyticity of γ_m in the variable t for all m (being the infinite power series $\gamma_{m,\infty}$ of (3.88) coincident with (3.86) in the whole domain $t \in]0;1]$).

To conclude the definition of this Taylor expansion, we must find a way to calculate the k$^{\text{th}}$ order derivative of γ_m at $t = 1$, for all m. If M is relatively low, that derivative can be simply computed by hand, but if we want M to be large or if we want to be able to tune that parameter dynamically to an arbitrary value, it is convenient to find an algorithmic way to compute the derivative. For that matter, define the function sequence $T_{k,m},\ m = 0, 1, \dots, N-1,\ k = 0, 1, 2, \dots$ as

$$T_{k,m}(z,t) = Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^{2k+2} |z(m)|^{2k} k! (-1)^k 2^{k-1}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\, |z(f)|^2 + 2|\hat{\bar{h}}_i(m)|^2\, |z(m)|^2\, (t-1) + c_2\right]^{k+1}} \tag{3.91}$$

for $m \neq 0$ and $m \neq N/2$, and as

$$T_{k,m}(z,t) = Nc_1 \sum_{q=0}^{\lfloor k/2 \rfloor} b_{k,q} t^{k-2q} \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^{2(k-q)+2} z(m)^{2(k-q)} (k-q)! (-1)^{k-q} 2^{k-q-1}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\, |z(f)|^2 + |\hat{\bar{h}}_i(m)|^2\, z^2(m)\, (t^2-1) + c_2\right]^{k-q+1}} \tag{3.92}$$

for $m = 0$ or $m = N/2$, where $b_{k,q}$ are the elements of the so-called triangle of the *Bessel numbers* [12], line k, column l. These elements satisfy the recurrence equation

$$\begin{aligned} b_{k,q} &= (k - 2q + 1)\, b_{k-1,q-1} + b_{k-1,q}, \quad q = 0, 1, \dots, \lfloor k/2 \rfloor,\ k = 1, 2, 3, \dots \\ b_{0,0} &= 1 \\ b_{k,-1} &= b_{k,\lfloor k/2 \rfloor + 1} = 0, \quad \forall k = 0, 1, 2, \dots. \end{aligned} \tag{3.93}$$

and have the closed-form expression

$$b_{k,q} = \binom{k}{2q} \frac{(2q)!}{q!\, 2^q}. \tag{3.94}$$

It can be seen, by inspection and direct computation of the derivative, that this sequence satisfies the properties

$$T_{0,m}(z,t) = \frac{1}{2}\, \gamma_m^2(z,t) \quad \text{and} \tag{3.95}$$

$$\frac{\partial T_{k,m}}{\partial t}(z,t) = T_{k+1,m}(z,t). \tag{3.96}$$

We will now get a recursive equation for the k$^{\text{th}}$ order derivative of γ_m using all orders smaller than k. To do this, we use the rule for the k$^{\text{th}}$ derivative of a product, proven in Appendix A, Theorem A.1, here stated for the product of γ_m with itself as

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k} \left[\gamma_m(x) \gamma_m(x)\right] = \sum_{n=0}^{k} \binom{k}{n} \frac{\mathrm{d}^n \gamma_m}{\mathrm{d}x^n}(x) \frac{\mathrm{d}^{k-n} \gamma_m}{\mathrm{d}x^{k-n}}(x). \tag{3.97}$$

If we use this rule with the function $T_{0,m}$, we obtain

$$\frac{\partial^k\, T_{0,m}}{\partial t^k}(z,t) = T_{k,m}(z,t) = \frac{1}{2} \frac{\partial\, \gamma_m^2}{\partial t}(z,t) = \frac{1}{2} \sum_{n=0}^{k} \binom{k}{n} \frac{\partial^n\, \gamma_m}{\partial t^n}(z,t) \frac{\partial^{k-n}\, \gamma_m}{\partial t^{k-n}}(z,t). \tag{3.98}$$

Solving for $\partial^k \gamma_m / \partial t^k$ (present in the first and last term of the summation), the required recurrence is obtained:

$$\frac{\partial^k \gamma_m}{\partial t^k}(z,t) = \gamma_m(z,t)^{-1}\left[T_{k,m}(z,t) - \frac{1}{2}\sum_{n=1}^{k-1}\binom{k}{n}\frac{\partial^n \gamma_m}{\partial t^n}(z,t)\frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,t)\right],$$
$$k = 1,2,3,\ldots, \tag{3.99}$$

where we define $\sum_{n=1}^{0} \overset{\text{def}}{=} 0$. Finally, Equation (3.99) can be further worked out due to the symmetry in its summation, being the result

$$\frac{\partial^k \gamma_m}{\partial t^k}(z,t) = \gamma_m(z,t)^{-1}\left[T_{k,m}(z,t) - \sum_{n=1}^{\frac{k-1}{2}}\binom{k}{n}\frac{\partial^n \gamma_m}{\partial t^n}(z,t)\frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,t)\right] \tag{3.100}$$

for k odd and

$$\gamma_m(z,t)^{-1}\left[T_{k,m}(z,t) - \sum_{n=1}^{\frac{k}{2}-1}\binom{k}{n}\frac{\partial^n \gamma_m}{\partial t^n}(z,t)\frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,t) - \binom{k-1}{k/2}\left(\frac{\partial^{k/2}\gamma_m}{\partial t^{k/2}}(z,t)\right)^2\right] \tag{3.101}$$

for k even.

To compress the signal using this Taylor expansion all that we need to do is to substitute $t = 1$ in (3.99) or (3.100) and (3.101), and in (3.91) or (3.92) for all m, using (3.87) and inserting the result in (3.90). In particular, with this substitution, the expressions (3.91) and (3.92) simplify to

$$T_{k,m}(z,1) = \begin{cases} Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\tilde{h}}_i(m)|^{2k+2}|z(m)|^{2k}k!(-1)^k 2^{k-1}}{\left[\sum_{f=0}^{N-1}|\hat{\tilde{h}}_i(f)|^2\,|z(f)|^2+c_2\right]^{k+1}} & \text{for } m \neq 0, N/2 \\ Nc_1 \sum_{i=0}^{P-1}\sum_{q=0}^{\lfloor k/2\rfloor} b_{k,q}\frac{|\hat{\tilde{h}}_i(m)|^{2(k-q)+2}z(m)^{2(k-q)}(k-q)!(-1)^{k-q}2^{k-q-1}}{\left[\sum_{f=0}^{N-1}|\hat{\tilde{h}}_i(f)|^2\,|z(f)|^2+c_2\right]^{k-q+1}} & \text{for } m = 0, N/2. \end{cases} \tag{3.102}$$

It is relevant to note that the calculation of the full compressor (the time-domain version $\tilde{F}$) boils down to applying Equation (3.64) adapted to this suboptimal compressor, i.e., we now do

$$\tilde{F}(x) = \frac{D_N^H}{\sqrt{N}}\,\tilde{G}(\sqrt{N}D_N\Lambda_w x). \tag{3.103}$$

As a last remark, note that when $N \to \infty$, due to the dominance expressed in Equation (3.79), the compressor gain degenerates into the 0^{th} order term of (3.89), so that asymptotically, the compressor just multiplies the input windowed signal in the frequency domain by the square-root of the inverse of the masking threshold, i.e.,

$$\Gamma_{m,M}(z)\big|_{N\to\infty} = \Gamma_{m,0}(z) = \gamma_m(z,1) = \hat{a}(z,m) \quad \text{and} \tag{3.104}$$
$$\tilde{G}_m(z)\big|_{N\to\infty} = \tilde{G}_{m,M}(z)\big|_{N\to\infty} = \sqrt{N}\,\hat{a}(z,m)\,z(m). \tag{3.105}$$

If you compare the compressor (3.105) with the normalization step (2.62), you will notice that asymptotically, the compressor does exactly the thing we wanted to avoid: normalize the input signal by the perceptual weights $\hat{\hat{a}}(z,m)$. Nevertheless, instead of transmitting these perceptual weights through the channel, at the receiver we now only have to apply the inverse of the compressor (3.105) (the expander); it is not necessary to use the weights at the receiver. How to calculate the inverse of the compressor will be the subject of Section 3.5.

3.4 Analysis of the Suboptimal Compressor

Now that we have developed a suboptimal compressor, both in its most complete form (Section 3.2) and performing a Taylor approximation on it (Section 3.3), it is of interest to analyze it in terms of its asymptotic behavior when $N \to \infty$. As our "benchmark criterion", the rate-loss (2.39), is given in terms of the Jacobian matrix of the compressor, that matrix will be calculated explicitly. Due to the lack of practical interest of Section 3.2 (derived from the previously explained computational burden of the corresponding compressor), this matrix will only be calculated for the compressor based on the Taylor expansion of Section 3.3. Note that, nevertheless, to obtain the exact expressions for the complete compressor, M can be set to infinity in the equations of this section due to the analyticity of γ_m of Equation (3.86) in t for all m.

After the calculation of the Jacobian Matrix of the suboptimal compressor, it will be time to analyze its behavior for $N \to \infty$, and we will conclude that it converges in weak norm to the optimal one of Equation (3.66). The asymptotic optimality of the scheme will then be naturally deduced from this behavior, being the final conclusion that the rate-loss (2.39) vanishes asymptotically.

3.4.1 Jacobian Matrix of the Compressor

Let us then calculate the components of the Jacobian matrix of (3.90). We will apply again the substitution trick relative to the hermitian symmetry of z of Equation (3.68) (see the surrounding text). Simple differentiation rules lead to[6]

$$\frac{\partial \tilde{G}_{m,M}}{\partial z_l}(z) = \sqrt{N} \begin{cases} \frac{\partial\,\Gamma_{m,M}}{\partial z_m}(z)\, z(m) + \Gamma_{m,M}(z) & \text{for } l = m \\ \frac{\partial\,\Gamma_{m,M}}{\partial z_l}(z)\, z(m) & \text{for } l \neq m \end{cases} \tag{3.106}$$

$$= \sqrt{N} \begin{cases} \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!} \left(\frac{\partial^{k+1}\,\gamma_m}{\partial z_m \partial t^k}(z,1)\, z(m) + \frac{\partial^k\,\gamma_m}{\partial t^k}(z,1) \right) & \text{for } l = m \\ \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!} \left(\frac{\partial^{k+1}\,\gamma_m}{\partial z_l \partial t^k}(z,1)\, z(m) \right) & \text{for } l \neq m. \end{cases} \tag{3.107}$$

[6] For $M = \infty$, due to the analyticity of $\gamma_m(z,t)$, the convergence of the Taylor series and of its term by term derivative is uniform so that we can differentiate term by term [5].

For the new term $\partial^{k+1}\gamma_m/(\partial z_l \partial t^k)$, we differentiate both sides of the recursion (3.99) (with $t = 1$) with respect to z_l, obtaining

$$\begin{aligned}\frac{\partial^{k+1}\gamma_m}{\partial z_l \partial t^k}(z,1) = &-\gamma_m(z,1)^{-2}\frac{\partial\gamma_m}{\partial z_l}(z,1)\left[T_{k,m}(z,1) - \frac{1}{2}\sum_{n=1}^{k-1}\binom{k}{n}\frac{\partial^n\gamma_m}{\partial t^n}(z,1)\frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,1)\right]\\ &+\gamma_m(z,1)^{-1}\left\{\frac{\partial T_{k,m}}{\partial z_l}(z,1) - \frac{1}{2}\sum_{n=1}^{k-1}\binom{k}{n}\left[\frac{\partial^{n+1}\gamma_m}{\partial z_l\partial t^n}(z,1)\frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,1)+\right.\right.\\ &\left.\left.+\frac{\partial^n\gamma_m}{\partial t^n}(z,1)\frac{\partial^{k-n+1}\gamma_m}{\partial z_l\partial t^{k-n}}(z,1)\right]\right\}.\end{aligned} \tag{3.108}$$

Reverse-substituting Equation (3.99) in (3.108), working out the result algebraically and taking advantage of the symmetry of the two terms in the last summation of (3.108), we can obtain a similar recursion to the one in (3.99), but now for the derivative $\partial^{k+1}\gamma_m/(\partial z_l \partial t^k)$.

$$\begin{aligned}\frac{\partial^{k+1}\gamma_m}{\partial z_l \partial t^k}(z,1) = &-\gamma_m(z,1)^{-1}\frac{\partial^k\gamma_m}{\partial t^k}(z,1)\frac{\partial\gamma_m}{\partial z_l}(z,1)+\\ &+\gamma_m(z,1)^{-1}\left[\frac{\partial T_{k,m}}{\partial z_l}(z,1) - \sum_{n=1}^{k-1}\binom{k}{n}\frac{\partial^{n+1}\gamma_m}{\partial z_l\partial t^n}(z,1)\frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,1)\right] \qquad (3.109)\\ =&\ \gamma_m(z,1)^{-1}\left[\frac{\partial T_{k,m}}{\partial z_l}(z,1) - \sum_{n=0}^{k-1}\binom{k}{n}\frac{\partial^{n+1}\gamma_m}{\partial z_l\partial t^n}(z,1)\frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,1)\right],\\ &k = 0,1,2,\ldots\end{aligned} \tag{3.110}$$

The validity of (3.110) for $k = 0$ (with $\sum_{n=0}^{-1} \overset{\text{def}}{=} 0$) comes from the differentiation of Equation (3.95). This recursion enables us to find the value for the k^{th} order term of the Taylor expansion of $\partial\gamma_m/\partial z_l$ around $t = 1$ (remember that the Schwarz' theorem holds) using all previous orders $0, 1, \ldots, k-1$, all outcomes of Equation (3.99) found so far (orders $1, 2, \ldots, k$) and a still to be calculated $\partial T_{k,m}/\partial z_l$.

As you can see from Equation (3.107), to compute the Jacobian matrix of the compressor, we multiply the result of (3.110) by $z(m)$ so that it will be more convenient to calculate directly the scaled version of (3.110) by $z(m)$. It is easy to see that recursion (3.110) is linear in the sense that if we insert a certain linear combination in the non-homogeneous term $\partial T_{k,m}/(\partial z_l)$, the solution we get is the same linear combination of each solution regarding each individual non-homogeneous term. We can thus state that if we define the scaled

$$D_{k,m,l}(z) \overset{\text{def}}{=} \frac{\partial^{k+1}\gamma_m}{\partial z_l\partial t^k}(z,1)\,z(m) \quad \text{and} \tag{3.111}$$

$$d_{k,m,l}(z) \overset{\text{def}}{=} \frac{\partial T_{k,m}}{\partial z_l}(z,1)\,z(m), \tag{3.112}$$

then (3.111) will be the solution of (3.110) if we insert (3.112) in the non-homogeneous term, i.e.,

$$D_{k,m,l}(z) = \gamma_m(z,1)^{-1}\left[d_{k,m,l}(z) - \sum_{n=0}^{k-1}\binom{k}{n} D_{n,m,l}(z)\frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,1)\right], \qquad k = 0,1,2,\ldots \tag{3.113}$$

The Jacobian matrix of compressor is then completely defined by equations (3.107), (3.113), (3.99) and its dependencies, and the scaled derivative of $T_{k,m}$ with respect to z_l, defined in (3.112), which will be calculated now. As previously stated, we will express $T_{k,m}$ (Equation (3.102)) as an analytic function using the hermitian symmetry of z, being thus our differentiation target

$$T_{k,m}(z,1) = \begin{cases} Nc_1 \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^{2k+2} z(m)^k z(N-m)^k k!(-1)^k 2^{k-1}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\, z(f)z(N-f)+c_2\right]^{k+1}} & \text{for } m \neq 0, N/2 \\ Nc_1 \sum_{q=0}^{\lfloor k/2\rfloor} b_{k,q} \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^{2(k-q)+2} z(m)^{2(k-q)} (k-q)!(-1)^{k-q} 2^{k-q-1}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\, z(f)(N-f)+c_2\right]^{k-q+1}} & \text{for } m = 0, N/2. \end{cases} \tag{3.114}$$

For $l \neq m$ and $l \neq N-m$, the computation of (3.112) yields

$$d_{k,m,l}(z)\Big|_{\substack{l\neq m,\\ N-m}} = z(m)z(l)^* \sum_{i=0}^{P-1} |\hat{\bar{h}}_i(l)|^2 Nc_1 \frac{|\hat{\bar{h}}_i(m)|^{2k+2}|z(m)|^{2k}(k+1)!(-1)^{k+1}2^k}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\,|z(f)|^2 + c_2\right]^{k+2}} \tag{3.115}$$

for $m \neq 0$ and $m \neq N/2$, being this also valid for $l = 0$ or $l = N/2$, and

$$\begin{aligned} d_{k,m,l}(z)\Big|_{\substack{l\neq m,\\ N-m}} &= z(m)z(l)^* \sum_{i=0}^{P-1} |\hat{\bar{h}}_i(l)|^2 \cdot \\ &\cdot Nc_1 \sum_{q=0}^{\lfloor k/2\rfloor} b_{k,q} \frac{|\hat{\bar{h}}_i(m)|^{2(k-q)+2} z(m)^{2(k-q)}(k-q+1)!(-1)^{k-q+1}2^{k-q}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\,|z(f)|^2 + c_2\right]^{k-q+2}} \end{aligned} \tag{3.116}$$

for $m = 0$ or $m = N/2$. For the case $l = m$ or $l = N-m$, a new term appears:

$$d_{k,m,l}(z)\Big|_{\substack{l=m,\\ N-m}} = d_{k,m,l}(z)\Big|_{\substack{l\neq m,\\ N-m}} + d_{k,m,l}(z)\Big|_{\substack{\text{new}\\ l=m,\\ N-m}}, \tag{3.117}$$

where

$$d_{k,m,l}(z)\Big|_{\substack{\text{new}\\ l=m,\\ N-m}} = \begin{cases} Nc_1 k \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^{2k+2}|z(m)|^{2k} k!(-1)^k 2^{k-1}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\,|z(f)|^2+c_2\right]^{k+1}} & \text{for } l = m \\ \mathrm{e}^{\mathrm{j}2\arg z(m)} Nc_1 k \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^{2k+2}|z(m)|^{2k} k!(-1)^k 2^{k-1}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2\,|z(f)|^2+c_2\right]^{k+1}} & \text{for } l = N-m \end{cases} \tag{3.118}$$

for $m \neq 0$ and $m \neq N/2$, where we define $\arg 0 \overset{\text{def}}{=} 0$ to handle de case $z(N-m) = z(m) = 0$, and

$$d_{k,m,l}(z)\Big|_{\substack{\text{new}\\ l=m}} = Nc_1 \sum_{q=0}^{\lfloor k/2 \rfloor} b_{k,q}(k-q) \sum_{i=0}^{P-1} \frac{|\hat{\bar{h}}_i(m)|^{2(k-q)+2} z(m)^{2(k-q)-1}(k-q)!(-1)^{k-q}2^{k-q}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2 \, |z(f)|^2 + c_2\right]^{k-q+1}} \tag{3.119}$$

for $m = 0$ or $m = N/2$. Notice that $N-m = m = l$ in this last case.

The Jacobian matrix of the full compressor $\tilde{F}_M$ comes directly from Equation (3.65) applied to the suboptimal compressor (although using a suboptimal compressor, we still use the scheme of Figure 3.1 with $\tilde{G}_M$ instead of G)

$$\tilde{F}'_M(x) = D_N^H \, \tilde{G}'_M(\sqrt{N} D_N \Lambda_w x) \, D_N \Lambda_w. \tag{3.120}$$

3.4.2 Rearranging the Jacobian matrix for Practical Feasibility

Although the Jacobian matrix of the compressor is completely defined now, if we would implement the results written above directly, we would have to compute (3.115) / (3.116) and also (3.117) with (3.118) / (3.119) for all m,$l = 0, 1, \ldots, N-1$ and $k = 0, 1, \ldots, M$, summing up a total of $N^2(M+1)$ executions. We would also have to run the recurrence (3.113) $N^2(M+1)$ times and finally apply (3.107) N^2 times (taking (3.111) into account), being thus the computational complexity $O(N^2M)$. The memory spent in the described process is also $O(N^2M)$ if we execute it exactly this way, or $O(N^2)$ if we do it component by component. If N is large (typically we work with the order of thousands) and if we do not need to calculate the Jacobian matrix explicitly, but we can leave it as the product of two matrices, then there is a much cheaper way to calculate those two matrices, both in terms of number of computations and memory.

To see this fact we have to do some calculations. First, take a look at equations (3.115), (3.116). The only dependence on l is on the terms $z^*(l)$ and $|\hat{\bar{h}}_i(l)|^2$, being the remaining part on the right a term similar to the summand of $T_{k,m}$ in i, Equation (3.102), not depending on l, but obviously depending on the summation variable i. Define then that last part as $\delta_{k,m,i}$,

$$\delta_{k,m,i}(z) = \begin{cases} Nc_1 \frac{|\hat{\bar{h}}_i(m)|^{2k+2} |z(m)|^{2k} (k+1)!(-1)^{k+1} 2^k}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2 \, |z(f)|^2 + c_2\right]^{k+2}} & \text{for } m \neq 0, N/2 \\ Nc_1 \sum_{q=0}^{\lfloor k/2 \rfloor} b_{k,q} \frac{|\hat{\bar{h}}_i(m)|^{2(k-q)+2} z(m)^{2(k-q)} (k-q+1)!(-1)^{k-q+1} 2^{k-q}}{\left[\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2 \, |z(f)|^2 + c_2\right]^{k-q+2}} & \text{for } m = 0, N/2. \end{cases} \tag{3.121}$$

We have thus

$$d_{k,m,l}(z)\Big|_{\substack{l \neq m,\\ N-m}} = z(m) z(l)^* \sum_{i=0}^{P-1} |\hat{\bar{h}}_i(l)|^2 \, \delta_{k,m,i}(z). \tag{3.122}$$

Due to the already mentioned linearity of the recursion (3.113), the solution of it, for $l \neq m$ and

$l \neq N-m$, when using (3.122), is of the form

$$D_{k,m,l}(z)\Big|_{\substack{l\neq m,\\ N-m}} = z(m)z(l)^* \sum_{i=0}^{P-1} |\hat{\bar{h}}_i(l)|^2 \, \Delta_{k,m,i}(z) \tag{3.123}$$

with $D_{k,m,i}$ coming out of the recursion

$$\Delta_{k,m,i}(z) = \gamma_m(z,1)^{-1} \left[\delta_{k,m,i}(z) - \sum_{n=0}^{k-1} \binom{k}{n} \Delta_{n,m,i}(z,1) \frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z) \right], \\ k = 0,1,2,\dots \tag{3.124}$$

for all $i = 0,1,\dots,P-1$. Due to the same reason, for $l = m$ or $l = N-m$, we can separately process the contributions of (3.117), so that

$$D_{k,m,l}(z)\Big|_{\substack{l=m,\\ N-m}} = D_{k,m,l}(z)\Big|_{\substack{l\neq m,\\ N-m}} + D_{k,m,l}(z)\Big|_{\substack{\text{new}\\ l=m,\\ N-m}} \tag{3.125}$$

with

$$D_{k,m,l}(z)\Big|_{\substack{\text{new}\\ l=m,\\ N-m}} = \gamma_m(z,1)^{-1} \left[d_{k,m,l}(z)\Big|_{\substack{\text{new}\\ l=m,\\ N-m}} - \sum_{n=0}^{k-1} \binom{k}{n} D_{n,m,l}(z)\Big|_{\substack{\text{new}\\ l=m,\\ N-m}} \frac{\partial^{k-n}\gamma_m}{\partial t^{k-n}}(z,1) \right], \\ k = 0,1,2,\dots \tag{3.126}$$

The branch of Equation (3.118) for $l = N-m$ is equal to the branch for $l = m$ up to a scaling factor $\mathrm{e}^{\mathrm{j}2\arg z(m)}$. The linearity of (3.126) delivers thus

$$D_{k,m,l}(z)\Big|_{\substack{\text{new}\\ l=N-m}} = \mathrm{e}^{\mathrm{j}2\arg z(m)} D_{k,m,l}(z)\Big|_{\substack{\text{new}\\ l=m}}. \tag{3.127}$$

Upon the substitution of the found results in (3.107) (using (3.111)), we finally arrive at an alternative expression for the elements of the Jacobian matrix of the compressor, namely, for $l \neq m$ and $l \neq N-m$

$$\frac{\partial \tilde{G}_{m,M}}{\partial z_l}(z)\Big|_{\substack{l\neq m,\\ N-m}} = \sqrt{N} \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!} \left(z(m)z(l)^* \sum_{i=0}^{P-1} |\hat{\bar{h}}_i(l)|^2 \, \Delta_{k,m,i}(z) \right) \tag{3.128}$$

$$= \sqrt{N} z(m)z(l)^* \sum_{i=0}^{P-1} |\hat{\bar{h}}_i(l)|^2 \left(\sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!} \Delta_{k,m,i}(z) \right), \tag{3.129}$$

and for $l = m$ or $l = N-m$

$$\frac{\partial \tilde{G}_{m,M}}{\partial z_l}(z)\Big|_{\substack{l=m,\\ N-m}} = \frac{\partial \tilde{G}_{m,M}}{\partial z_l}(z)\Big|_{\substack{l\neq m,\\ N-m}} + \frac{\partial \tilde{G}_{m,M}}{\partial z_l}(z)\Big|_{\substack{\text{new}\\ l=m,\\ N-m}} \tag{3.130}$$

with

$$\frac{\partial \tilde{G}_{m,M}}{\partial z_l}(z)\Big|_{\substack{\text{new}\\ l=m,\\ N-m}} = \sqrt{N}\begin{cases} \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!}\left(D_{k,m,l}(z)\big|_{\substack{\text{new}\\ l=m}} + \frac{\partial^k \gamma_m}{\partial t^k}(z,1)\right) & \text{for } l=m \\ \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!}\,\mathrm{e}^{\mathrm{j}2\arg z(m)} D_{k,m,l}(z)\big|_{\substack{\text{new}\\ l=m}} & \text{for } l=N-m. \end{cases} \tag{3.131}$$

In matrix notation, define the N-by-P tall matrices (P is usually much smaller than N) A and H by

$$[A(z)]_{m,i} = z(m)\left(\sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!}\Delta_{k,m,i}(z)\right) \quad \text{and} \tag{3.132}$$

$$[H(z)]_{m,i} = z(m)|\hat{\bar{h}}_i(m)|^2, \tag{3.133}$$

$$m = 0,1,2,\ldots,N-1,\ i=0,1,\ldots,P-1, \tag{3.134}$$

respectively. Furthermore, define an N-by-N matrix V as

$$V(z) = \Lambda_{v_f(z)} + \Lambda_{v_b(z)} D_N^2 \tag{3.135}$$

with

$$[v_f(z)]_m = \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!}\left(D_{k,m,l}(z)\big|_{\substack{\text{new}\\ l=m}} + \frac{\partial^k \gamma_m}{\partial t^k}(z,1)\right) \tag{3.136}$$

$$[v_b(z)]_m = \begin{cases} \sum_{k=0}^{M} \frac{(-1)^k}{(k+1)!}\,\mathrm{e}^{\mathrm{j}2\arg z(m)} D_{k,m,l}(z)\big|_{\substack{\text{new}\\ l=m}} & \text{for } m \neq 0, N/2 \\ 0 & \text{for } m = 0, N/2 \end{cases}, \tag{3.137}$$

$$m = 0,1,2,\ldots,N-1.$$

We name a matrix with the format of (3.135) as a *cross-diagonal* matrix. This kind of matrices has non-zero entries only on the main diagonal and on the elements of D_N^2 with value 1 (cf. matrix (3.60)). Using these newly defined matrices, the Jacobian matrix of the compressor in the frequency domain, using the approximation by a Taylor expansion, is given by

$$\tilde{G}'_M(z) = \sqrt{N}(V(z) + A(z)H(z)^{\mathrm{H}}). \tag{3.138}$$

It is relevant to note here that, due to the fact that the non-homogeneous terms of the recursions (3.99), (3.124) and (3.126) for $l=m$, $T_{k,m}$, $\delta_{k,m,i}$ and $(d_{k,m,l})_{\substack{\text{new}\\ l=m}}$, respectively, are real and have their dependence on m only on magnitudes of hermitian symmetric signals (z and $\hat{\bar{h}}_i$), $\Delta_{k,m,i}$ and $(D_{k,m,l})_{\substack{\text{new}\\ l=m}}$ turn out to be real and symmetric in m, with the consequence that the matrices A and H have hermitian symmetric columns (i.e., $D_N^2 A = A^*$ and the same for H), v_f is real and symmetric ($v_f = v_f^*$ and $D_N^2 v_f = v_f$), and v_b is hermitian symmetric ($D_N^2 v_b = v_b^*$).

Equation (3.138) expresses the Jacobian matrix of the compressor (in its Taylor expansion form) as the sum of a large, sparse N-by-N cross-diagonal matrix, which is in principle full-rank, by a low-rank matrix AH^{H}. If we do not need to calculate $\tilde{G}'_M$ explicitly, but only v_f, v_b, A and H, then, to calculate A we only need to run (3.124) $N(M+1)P$ times and (3.132) NP times (with $M+1$ terms in the summation of each run). For v_f and v_b, the computation of $(D_{k,m,l})^{\text{new}}_{l=m}$ has complexity $O(NM)$ and the computation of (3.126) / (3.127), $O(M)$, so that the overall complexity of the scheme is $O(NMP)$. In terms of the memory usage, we have again $O(NP)$ for a component-wise computation and $O(NMP)$ for a computation of the equations step-by-step for all components at the same time. As P is usually on the order of dozens (we used $P = 60$ in practice) and N is on the order of thousands ($N = 1024$ is a reasonable value), there is quite a performance gain. Furthermore, with this optimization, the computation time and memory usage vary linearly with N and not quadratically, due to the number of filters P remaining constant with varying N.

3.4.3 Asymptotic Optimality of the Compressor

We have now derived expressions which enable the computation of the derivative of the compressor in a recursive form and worked them out into a form which sped up its execution time and amount of memory spent. But what about the performance of the companding scheme associated to this compressor with respect to the rate-distortion function? This performance is measured in terms of the rate-loss of Equation (2.39) and, indirectly, through the proximity of the jacobian matrix $\tilde{G}'_M$ to the ideal one of Equation (3.66). We will prove in this subsection that their distance, in terms of the Hilbert-Schmidt norm, disappears with N asymptotically large, making the rate-loss vanish under the same conditions.

As previously explained (cf. equations (3.104) and (3.105) and corresponding text), asymptotically, the 0^{th} order term of the compressor gain is dominant, so that the compressor boils down to multiplying the input signal (in the frequency domain) by the square-root of the inverse of the masking threshold. The original compressor of Section 3.2 is thus asymptotically the same as the one corresponding to the Taylor expansion of Section 3.3 for any M; they both collapse to the one with $M = 0$. It is then intuitive, also for large N, that the derivative of the compressor may be well approximated by the derivative of the asymptotic compressor, i.e., by the 0^{th} order terms of the equations deduced in Subsection 3.4.1. To see that this is indeed the case, let us check the asymptotic behavior of those equations. Using the same approach as the one we used to derive equations (3.83) to (3.85), it is true that

$$\lim_{N\to\infty} \frac{|\hat{\bar{h}}_i(m)|^2 \, Z(m) Z(l)^*}{\sum_{f=0}^{N-1} |\hat{\bar{h}}_i(f)|^2 \, |Z(f)|^2 + c_2} = 0 \tag{3.139}$$

in probability, for all l,$m = 0,1,\ldots,N-1$ with a convergence rate of $O(1/N)$, i.e., the left hand side becomes deterministic and equal to 0 for all m when $N \to \infty$ and on average it decreases at a rate equal to the one of the sequence $1/N$. The main differences towards the previous deduction are on the term c_2 and on the usage of $Z(l)^*$ instead of $Z(m)^*$. If we assume L proportional to N (for large N, we even set usually $L = N$), then c_2 is proportional to N^2 (cf. Subsection 2.3.2) so that the effect of it on the previous deduction is just a summation of a constant (the constant c_2') in the denominator of the right hand side of Equation (3.82). Furthermore, the amplitude of $Z(l)/\sqrt{N}$ has the same behavior (in terms of its expected value and variance) as the one of $Z(m)/\sqrt{N}$ with varying N (for $l \neq m$), as it belongs to the same (square-root of the) periodogram, but simply calculated on another frequency value.

We can also formulate a similar convergence statement for a similar sequence of random variables than the one of Equation (3.139), but without signal on the numerator. We have in that case

$$\lim_{N\to\infty} \frac{|\hat{\hat{h}}_i(m)|^2}{\sum_{f=0}^{N-1} |\hat{\hat{h}}_i(f)|^2 \, |Z(f)|^2 + c_2} = 0. \tag{3.140}$$

Here,

$$\frac{1}{N^2}\left(\sum_{f=0}^{N-1} |\hat{\hat{h}}_i(f)|^2 \, |Z(f)|^2 + c_2\right) \tag{3.141}$$

converges in probability to a non-zero value when $N \to \infty$ (to the power of X plus a constant), so that, due to the continuous mapping theorem [16], (3.140) converges to 0 at a rate $O(1/N^2)$.

Due to the equivalence between convergence in probability and convergence in distribution when the limit is a deterministic constant [16], we can use Slutsky's theorem [26] and limits (3.139) and (3.140) to state that the variable $d_{k,m,l}$ of equations (3.115) and (3.116) converges in probability to 0 when $N \to \infty$ at a rate $O(1/N^{k+2})$, $k \geq 0$. Using the continuous mapping theorem again, $\gamma_m(t, 1)$ of Equation (3.86) does also converge to 0, but at a rate $O(1/\sqrt{N})$. Furthermore, in Equation (3.113), the lowest order term of $\partial^{k-n}\gamma_m/\partial t^{k-n}$ that we use is for $k-n = 1$, and as $T_{k,m}$ of Equation (3.102) vanishes with $O(1/N^{k+1})$, the former function, expressed in Equation (3.99), vanishes at a rate at least $O(1/N^{1,5})$, i.e., it vanishes with $O(1/N^{r/2})$, $r \geq 3$. Consequently, the result of Equation (3.113) also vanishes with $O(1/N^{r/2})$, $r \geq 3$, and the slowest rate of convergence occurs with $k = 0$ (for $k = 0$ we have thus a rate $O(1/N^{3/2})$ and for example for $k = 1$ a rate $O(1/N^{5/2})$). We can finally conclude that the dominant term in the elements of the Jacobian matrix (3.107) is $\partial^0\gamma_m/\partial t^0 = \gamma_m$, which occurs for $l = m$, and which decreases at a rate $O(1/\sqrt{N})$. This term is followed by $\partial\gamma_m/\partial z_l \, z(m) = D_{0,m,l}$ (both for $l = m$ and $l \neq m$), decreasing at a rate $O(1/N^{3/2})$. For $l \neq m$, the term $D_{0,m,l}$ is now the most slowly falling. The terms that were pointed out are exactly the ones of the Jacobian matrix of the compressor for $M = 0$.

The last paragraph should have already given some intuition why the compressor is asymptotically optimal: the terms $D_{k,m,l}$ all fall at a rate at least $O(1/N^{3/2})$, whereas the term γ_m falls much more slowly, at a pace $O(1/\sqrt{N})$. The former terms become thus negligible for $N \to \infty$ with respect to the latter one, and if we have the latter term alone, we fulfill the optimality conditions of Equation (3.69). More formally, let us try to calculate the Hilbert-Schmidt distance from $G'(z)^{-1}\tilde{G}'_M(z)$ to the identity matrix. First, look at the Jacobian matrix of the suboptimal compressor, for $M = 0$. As already explained, the neglectfulness of $D_{k,m,l}$ for $k > 0$ makes the exact derivative for $M \neq 0$ arbitrarily near to this one when $N \to \infty$. By simple substitution of equations (3.115), (3.116) and (3.86) in Equation (3.113) and then (3.107), using further the definition (3.111) and the identity (3.87), we get

$$\frac{\partial \tilde{G}_{m,M}}{\partial z_l}(z)\Big|_{N\to\infty} = \frac{\partial \tilde{G}_{m,0}}{\partial z_l}(z) = \sqrt{N}\begin{cases} \frac{d_{0,m,l}(z)}{\gamma_m(z,1)} + \gamma_m(z,1) & \text{for } l = m \\ \frac{d_{0,m,l}(z)}{\gamma_m(z,1)} & \text{for } l \neq m \end{cases} \tag{3.142}$$

$$= \sqrt{N}\gamma_m(z,1)\left(\frac{d_{0,m,l}(z)}{\gamma_m(z,1)^2} + \mathrm{I}_{ml}\right) \tag{3.143}$$

$$= \sqrt{N}\hat{\hat{a}}(z,m)\left(\frac{-Nc_1\sum_{i=0}^{P-1}\frac{|\hat{\hat{h}}_i(m)|^2|\hat{\hat{h}}_i(l)|^2\, z(m)z(l)^*}{\left[\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|z(f)|^2+c_2\right]^2}}{Nc_1\sum_{i=0}^{P-1}\frac{|\hat{\hat{h}}_i(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|z(f)|^2+c_2}} + \mathrm{I}_{m,l}\right), \tag{3.144}$$

where $\mathrm{I}_{m,l}$ is the Kronecker delta at $(m{,}l)$, i.e., the element $(m{,}l)$ of the identity matrix. Note that equations (3.118) and (3.119) are 0 in this case ($k = 0$). From (3.66), it is then obvious that

$$\left[G'(Z)^{-1}\tilde{G}'_M(Z)\Big|_{N\to\infty}\right]_{m,l} = \frac{-\sum_{i=0}^{P-1}\frac{|\hat{\hat{h}}_i(m)|^2|\hat{\hat{h}}_i(l)|^2\, Z(m)Z(l)^*}{\left[\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|Z(f)|^2+c_2\right]^2}}{\sum_{i=0}^{P-1}\frac{|\hat{\hat{h}}_i(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|Z(f)|^2+c_2}} + \mathrm{I}_{m,l} \tag{3.145}$$

and, from definition (3.35), now for the field of the complex numbers (i.e., using squared magnitudes),

$$\left\| G'(Z)^{-1}\tilde{G}'_M(Z)\Big|_{N\to\infty} - \mathrm{I}\right\|_{\mathrm{HS}}^2 = \frac{1}{N}\sum_{m,l=0}^{N-1}\left|\frac{\sum_{i=0}^{P-1}\frac{|\hat{\hat{h}}_i(m)|^2|\hat{\hat{h}}_i(l)|^2\, Z(m)Z(l)^*}{\left[\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|Z(f)|^2+c_2\right]^2}}{\sum_{i=0}^{P-1}\frac{|\hat{\hat{h}}_i(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|Z(f)|^2+c_2}}\right|^2. \tag{3.146}$$

We can upper bound this last equation by taking the maximum of the summand:

$$\left\| G'(Z)^{-1}\tilde{G}'_M(Z)\Big|_{N\to\infty} - \mathrm{I}\right\|_{\mathrm{HS}}^2 \leq N\max_{m,l=0}^{N-1}\left|\frac{\sum_{i=0}^{P-1}\frac{|\hat{\hat{h}}_i(m)|^2|\hat{\hat{h}}_i(l)|^2\, Z(m)Z(l)^*}{\left[\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|Z(f)|^2+c_2\right]^2}}{\sum_{i=0}^{P-1}\frac{|\hat{\hat{h}}_i(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|Z(f)|^2+c_2}}\right|^2. \tag{3.147}$$

From limits (3.139) and (3.140) and Slutsky's theorem, the term in the numerator of the right hand side of (3.147) (inside the squared magnitude) converges in probability to 0 for all $m{,}l = 0{,}1{,}\ldots{,}N-1$ at a rate $O(1/N^3)$ when $N \to \infty$. The same happens to the denominator, but at a rate $O(1/N^2)$. By dominance of the numerator, the quotient also converges to 0, namely at a rate

$O(1/N)$. The squared magnitude of the fraction converges thus at rate $O(1/N^2)$ and the complete expression converges to 0 at a rate $O(1/N)$. We have thus

$$\lim_{N\to\infty} \left\| \left. G'(Z)^{-1}\tilde{G}'_M(Z)\right|_{N\to\infty} - \mathrm{I} \right\|_{\mathrm{HS}} = 0 \tag{3.148}$$

in probability, at a rate $O(1/\sqrt{N})$, i.e., $G'(z)^{-1}\tilde{G}'_M(z)$ is asymptotically equivalent to the identity matrix, in the sense explained in [25]. We will denote this asymptotic behavior by

$$G'(Z)^{-1}\tilde{G}'_M(Z) \overset{N\to\infty}{\to} \mathrm{I}. \tag{3.149}$$

As we will see in Section 3.5, the Jacobian matrix of the compressor with the 0^{th} order Taylor expansion of the integrand $\gamma_m(z,t)$ $(M = 0)$, whose elements are written in (3.144), is invertible even with finite N (with $N \to \infty$ this statement is trivial, as we fulfill (3.66) and $\hat{\tilde{a}}(x, f) > 0$, $\forall f$). If we further admit that this happens for all other orders $M = 1,2,\ldots,\infty$, we can finally use simple matrix convergence properties deduced in [25, Theorem 2.1] and the invariance of the Hilbert-Schmidt norm upon hermitian transposition of matrices to obtain

$$\tilde{G}'_M(Z) \overset{N\to\infty}{\to} G'(Z), \tag{3.150}$$

$$\tilde{G}'_M(Z)^{\mathrm{H}} \overset{N\to\infty}{\to} G'(Z)^{\mathrm{H}} \quad \text{and} \tag{3.151}$$

$$\tilde{G}'_M(Z)^{-1}\tilde{G}'_M(Z)^{-\mathrm{H}}G'(Z)^{\mathrm{H}}G'(Z) \overset{N\to\infty}{\to} \mathrm{I}. \tag{3.152}$$

From equations (3.65), (3.120) and Equation (2.38), we can see that

$$\operatorname{tr}[\tilde{M}(X)^{-1}M(X)] = \operatorname{tr}\left[\left(\tilde{F}'(X)^{\mathrm{T}}\tilde{F}'(X)\right)^{-1} F'(X)^{\mathrm{T}}F'(X)\right] \tag{3.153}$$

$$= \operatorname{tr}\left[\tilde{G}'_M(Z)^{-1}\tilde{G}'_M(Z)^{-\mathrm{H}}G'(Z)^{\mathrm{H}}G'(Z)\right] \tag{3.154}$$

and

$$\frac{\det \tilde{M}(X)}{\det M(X)} = \frac{1}{\det[\tilde{M}(X)^{-1}M(X)]} = \frac{1}{\det[\tilde{G}'_M(Z)^{-1}\tilde{G}'_M(Z)^{-\mathrm{H}}G'(Z)^{\mathrm{H}}G'(Z)]} \tag{3.155}$$

so that, due to the asymptotic equivalent distribution of the eigenvalues of the matrices on the left and right hand side of (3.152), we can use (3.56) with $\Psi(v) = v$ for the trace and $\Psi(v) = \log_2(v)$ for the determinant to state that, asymptotically, the rate-loss (2.39) vanishes and the companding scheme is optimal.

3.5 Expander

After having derived a suboptimal compressor and having analyzed it in terms of its Jacobian matrix and its asymptotic behavior with increasing vector dimension, we would like to complete the whole chain of the companding scheme depicted in Figure 1.2 by building an expander which implements the inverse function of the suboptimal compressor developed in sections 3.2 and 3.3.

First, we must first prove that such an inverse function exists, at least locally (when the domain of F is a small neighborhood of a realization x of the input signal X). That proof will only be done for a Taylor order of the compressor $M = 0$ due to the complexity of the equations for $M > 0$. Nevertheless, as we will see in the simulations (Section 4.3) the compressor with this order of the Taylor expansion is the most interesting one, since the contribution of the 0^{th} order term to the gain Γ_m of Equation (3.89) is dominant. Following the proof, we build an expander based on numerical methods, using as an initial estimate the signal from the previous audio frame. Finally, a memory optimization of the expander is done, so that it can run on very large vector dimensions; in practice, values as large as $N = 65536$ could be achieved.

3.5.1 Invertibility of the Compressor

Obviously, due to the one-to-one relation between the whole compressor $\tilde{F}$ and the compressor in the frequency domain $\tilde{G}$, depicted in Figure 3.1 (adapted to the suboptimal compressor), $\tilde{F}$ is invertible if and only if $\tilde{G}$ is invertible as well. Indeed, using Equation (3.103), we can get the equivalence

$$\xi = \tilde{F}(y) = \frac{D_N^{\mathrm{H}}}{\sqrt{N}} \tilde{G}(\sqrt{N} D_N \Lambda_w y) \iff y = \tilde{F}^{-1}(\xi) = \Lambda_w^{-1} \frac{D_N^{\mathrm{H}}}{\sqrt{N}} \tilde{G}^{-1}(\sqrt{N} D_N \xi), \tag{3.156}$$

so that if we know that $\tilde{G}$ is invertible, then we conclude that $\tilde{F}$ is invertible as well. By rearranging the left hand side of (3.156) so that we have $\tilde{G}$ in terms of $\tilde{F}$, we can also construct a similar equivalence, concluding that if $\tilde{F}$ is invertible, so is $\tilde{G}$.

We consequently only have to invert $\tilde{G}$ at point $\hat{\xi} = \sqrt{N} D_N \xi$ to find out the inverse of $\tilde{F}$. Due to the high mathematical complexity of the equations of sections 3.2 and 3.3, no proof could be found for the invertibility of $\tilde{G}$ for an order M of the Taylor expansion greater than 0. Nevertheless, as it will be shown in the simulations, Section 4.3, the compressor with order $M = 0$ is already an excellent approximation of the compressor with $M = \infty$, so that $M = 0$ is the most interesting (and the least computationally expensive) case to study. Let us then try to get an expression for the determinant of $\tilde{G}'_M(z)$ for $M = 0$ (or equivalently, for $N \to \infty$) to check whether this matrix is invertible or not. To do that, we first write (3.144) in matrix form, getting an expression similar to (3.138):

$$\tilde{G}'(z)\Big|_{N\to\infty} = \tilde{G}'_M(z)\Big|_{M=0} = \Lambda_{\sqrt{N}\hat{a}(z)} \left(\mathrm{I} + [-\Lambda_{\hat{\hat{a}}(z)^{-2}} A'(z)] A'(z)^{\mathrm{H}}\right) \tag{3.157}$$

with

$$\begin{aligned}[A'(z)]_{m,i} &= \frac{\sqrt{Nc_1}\,|\hat{\hat{h}}_i(m)|^2 z(m)}{\sum_{f=0}^{N-1} |\hat{\hat{h}}_i(f)|^2\,|z(f)|^2 + c_2},\\ &\qquad m = 0,1,2,\dots,N-1,\ i = 0,1,\dots,P-1.\end{aligned} \tag{3.158}$$

Using the *Matrix Determinant Lemma* [27],

$$\begin{aligned}\det\left[\tilde{G}'_M(z)\Big|_{M=0}\right] &= \det[\Lambda_{\sqrt{N}\hat{\mathrm{a}}(z)}]\det\left[\mathrm{I} + [-\Lambda_{\hat{\mathrm{a}}(z)^{-2}}A'(z)]A'(z)^{\mathrm{H}}\right] && (3.159)\\ &= \det[\Lambda_{\sqrt{N}\hat{\mathrm{a}}(z)}]\det\left[\mathrm{I} - A'(z)^{\mathrm{H}}\Lambda_{\hat{\mathrm{a}}(z)^{-2}}A'(z)\right], && (3.160)\end{aligned}$$

which means that $\tilde{G}'_M$ (with $M = 0$) is invertible if and only $\Lambda_{\sqrt{N}\hat{\mathrm{a}}}$ and $\mathrm{I} - {A'}^{\mathrm{H}}\Lambda_{\hat{\mathrm{a}}^{-2}}A'$ are both invertible. This is obviously valid for the former matrix, since $\hat{\mathrm{a}}(z,m) > 0,\ m = 0,1,\dots,N-1,\forall z$ hermitian symmetric. For the latter matrix, this validity is not obvious at first sight; we will use the *Levy-Desplanques theorem* to prove that this is indeed the case. If we express each component (i,j) of the matrix explicitly, we get

$$\begin{aligned}\left[\mathrm{I} - A'(z)^{\mathrm{H}}\Lambda_{\hat{\mathrm{a}}(z)^{-2}}A'(z)\right]_{i,j} &= \mathrm{I}_{i,j} - \sum_{m=0}^{N-1} \frac{\frac{|\hat{\hat{h}}_i(m)|^2|\hat{\hat{h}}_j(m)|^2|z(m)|^2}{\left[\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|z(f)|^2+c_2\right]\left[\sum_{f=0}^{N-1}|\hat{\hat{h}}_j(f)|^2\,|z(f)|^2+c_2\right]}}{\sum_{q=0}^{P-1}\frac{|\hat{\hat{h}}_q(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_q(f)|^2\,|z(f)|^2+c_2}},\\ &\qquad i,j = 0,1,\dots,P-1.\end{aligned} \tag{3.161}$$

We can assume that $z \neq 0$, since even in silent periods of audio frames there is always noise. The sum on the right hand side of (3.161), representing the components of ${A'}^{\mathrm{H}}\Lambda_{\hat{\mathrm{a}}^{-2}}A'$, is then strictly positive as its terms are derived from squared magnitudes (where at least one term is non-zero). Furthermore, if we sum the columns of the mentioned matrix, we get

$$\begin{aligned}\sum_{j=0}^{P-1}\left[A'(z)^{\mathrm{H}}\Lambda_{\hat{\mathrm{a}}(z)^{-2}}A'(z)\right]_{i,j} &= \sum_{j=0}^{P-1}\sum_{m=0}^{N-1} \frac{\frac{|\hat{\hat{h}}_i(m)|^2|\hat{\hat{h}}_j(m)|^2|z(m)|^2}{\left[\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|z(f)|^2+c_2\right]\left[\sum_{f=0}^{N-1}|\hat{\hat{h}}_j(f)|^2\,|z(f)|^2+c_2\right]}}{\sum_{q=0}^{P-1}\frac{|\hat{\hat{h}}_q(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_q(f)|^2\,|z(f)|^2+c_2}} && (3.162)\\ &= \sum_{m=0}^{N-1}\frac{\frac{|\hat{\hat{h}}_i(m)|^2|z(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|z(f)|^2+c_2}\sum_{j=0}^{P-1}\frac{|\hat{\hat{h}}_j(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_j(f)|^2\,|z(f)|^2+c_2}}{\sum_{q=0}^{P-1}\frac{|\hat{\hat{h}}_q(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_q(f)|^2\,|z(f)|^2+c_2}} && (3.163)\\ &= \frac{\sum_{m=0}^{N-1}|\hat{\hat{h}}_i(m)|^2|z(m)|^2}{\sum_{f=0}^{N-1}|\hat{\hat{h}}_i(f)|^2\,|z(f)|^2+c_2} < 1 && (3.164)\end{aligned}$$

since $c_2 > 0$. As each term of that matrix is positive, each term must lie between 0 and 1.

Consequently, the magnitude of the $(i,i)^{\text{th}}$ term of $\mathrm{I} - {A'}^{\mathrm{H}}\Lambda_{\hat{\bar{a}}^{-2}}A'$ is equal to the term itself,

$$\left|\left[\mathrm{I} - A'(z)^{\mathrm{H}}\Lambda_{\hat{\bar{a}}(z)^{-2}}A'(z)\right]_{i,i}\right| = 1 - \left[A'(z)^{\mathrm{H}}\Lambda_{\hat{\bar{a}}(z)^{-2}}A'(z)\right]_{i,i}, \tag{3.165}$$

and the magnitude of the $(i,j)^{\text{th}}$ term, for $j \neq i$ is equal to the symmetric of the term,

$$\left|\left[\mathrm{I} - A'(z)^{\mathrm{H}}\Lambda_{\hat{\bar{a}}(z)^{-2}}A'(z)\right]_{i,j}\right| = \left[A'(z)^{\mathrm{H}}\Lambda_{\hat{\bar{a}}(z)^{-2}}A'(z)\right]_{i,j} \tag{3.166}$$

If we subtract the sum of the elements of (3.166) from (3.165), we have thus

$$\left|\left[\mathrm{I} - A'(z)^{\mathrm{H}}\Lambda_{\hat{\bar{a}}(z)^{-2}}A'(z)\right]_{i,i}\right| - \sum_{\substack{j=0\\ j\neq i}}^{P-1}\left|\left[\mathrm{I} - A'(z)^{\mathrm{H}}\Lambda_{\hat{\bar{a}}(z)^{-2}}A'(z)\right]_{i,j}\right| = 1 - \sum_{j=0}^{P-1}\left[A'(z)^{\mathrm{H}}\Lambda_{\hat{\bar{a}}(z)^{-2}}A'(z)\right]_{i,j} \tag{3.167}$$

and

$$1 - \sum_{j=0}^{P-1}\left[A'(z)^{\mathrm{H}}\Lambda_{\hat{\bar{a}}(z)^{-2}}A'(z)\right]_{i,j} = 1 - \frac{\sum_{m=0}^{N-1}|\hat{\bar{h}}_i(m)|^2|z(m)|^2}{\sum_{f=0}^{N-1}|\hat{\bar{h}}_i(f)|^2\,|z(f)|^2 + c_2} > 0. \tag{3.168}$$

Equations (3.167) and (3.168) tell us that $\mathrm{I} - {A'}^{\mathrm{H}}\Lambda_{\hat{\bar{a}}^{-2}}A'$ is strictly diagonally dominant. By the *Gershgorin circle theorem* [29], each eigenvalue of this matrix lies on an interval[7] not containing the origin, with the consequence that the matrix is invertible. This is known as the Levy-Desplanques theorem [29].

3.5.2 A Numerical Expander

After some calculations, we arrived finally at the conclusion that $\tilde{G}'_M$ is invertible when $M = 0$, for all N. By the implicit function theorem in complex analysis [54], there is a local neighborhood of the function $\tilde{G}_0$ where it is invertible (i.e., where $\tilde{G}_0^{-1}$ exists), when defined on that domain. We are thus sure that, for $M = 0$, if we use an adequate numerical method to solve the equation

$$\tilde{G}(z) = \hat{\xi} \tag{3.169}$$

with a sufficiently close initial estimate $z^{(0)}$, the solution exists and the method will converge to it. We assume that this also happens for $M = 1, 2, \ldots, \infty$. For the inversion of the complete compressor function $\tilde{F}(x)$, we only have to use Equation (3.156).

For the first iteration of the method, we propose to rearrange Equation (3.77) as

$$z(m) = \frac{\hat{\xi}}{\sqrt{N}\,\Gamma_m(z)} \tag{3.170}$$

[7] The eigenvalues are real due to the symmetry of the matrix.

and perform a fixed-point iteration of it, i.e., build a second estimate $z^{(1)}$ with

$$z^{(1)}(m) = \frac{\hat{\xi}}{\sqrt{N}\,\Gamma_m(z^{(0)})}, \quad m = 0, 1, \ldots, N-1. \tag{3.171}$$

This first iteration can be motivated as follows. In the first place, remember that $\Gamma(z)$, Equation (3.89), was obtained using a recursive equation (Equation (3.99)), where the non-homogeneous term, Equation (3.102), was not dependent on the phase of the components of z, but only on their magnitude. The execution of iteration (3.171) has thus the advantage that it wipes out phase differences between the initial estimate $z^{(0)}$ and the final desired value z of (3.169). In other words, if $z^{(0)}$ is very close to z up to phase differences, the second estimate $z^{(1)}$ will be an excellent initial estimate for the next numerical method, which will be used to "fine-tune" the obtained vector $z^{(1)}$. The same argument can be applied if $z^{(0)}$ has a similar masking threshold $\hat{a}^{-2}$ than the one of z. Remember that the 0^{th} order term of Γ_m is exactly the square-root of the inverse of the masking threshold and that for high vector dimension, only this term matters (Equation (3.104)). For similar masking thresholds we get thus similar $\Gamma(z)$ vectors.

For the initial estimate $z^{(0)}$, we propose to use the vector z obtained from running the numerical methods described in this section on the last audio frame. With this choice, the masking threshold will not have changed much between the last and current frames due to the stationarity of typical audio signals in the order of the dozens of milliseconds. By using Equation (3.171) in this way, we thereby get a good initial estimate for fine-tuning.

For the fine-tuning process, we could use for example the Newton's method [44] for $\tilde{G}(z) - \hat{\xi}$,

$$z^{(n+1)} = z^{(n)} - \tilde{G}'(z^{(n)})^{-1}\left(\tilde{G}(z^{(n)}) - \hat{\xi}\right), \quad n = 1, 2, 3, \ldots, \tag{3.172}$$

with the Jacobian matrix $\tilde{G}'^{-1}$ derived in subsections 3.4.1 and 3.4.2, until successive estimates are equal up to numerical noise, i.e., until

$$\frac{\|z^{(n+1)} - z^{(n)}\|}{\sqrt{N}} < \epsilon, \tag{3.173}$$

where ϵ is the numerical unit round-off of the machine executing the algorithm.

Although this process is feasible, calculating the Jacobian matrix was found be, in practice, computationally expensive. For speeding up this process we thus propose to use a quasi-Newton method, also of the form

$$z^{(n+1)} = z^{(n)} - \tilde{J}_n^{-1}\left(\tilde{G}(z^{(n)}) - \hat{\xi}\right), \quad n = 1, 2, 3, \ldots \tag{3.174}$$

but where now $\tilde{J}_n$ is an approximation to $\tilde{G}'(z^{(n)})$, instead of its real version. One possibility to determine $\tilde{J}_n$ is to use the equivalent of the secant method in multiple dimensions, the *Broyden's* *method* [44]. In this method, $\tilde{J}_n$ is chosen such that it satisfies the secant equation

$$\tilde{J}_n\left(z^{(n)} - z^{(n-1)}\right) = \tilde{G}(z^{(n)}) - \tilde{G}(z^{(n-1)}). \tag{3.175}$$

As this equation is underdetermined for $N > 1$ (it has N^2 unknowns, namely the components of $\tilde{J}_n$ and N equations), we have to pose an additional constraint to get a specific solution $\tilde{J}_n$. Broyden chose the solution $\tilde{J}_n$ that minimized the Frobenius distance to the matrix of the previous iteration $\tilde{J}_{n-1}$, i.e., $\tilde{J}_n$ fulfills

$$\tilde{J}_n = \operatorname{argmin}_{\tilde{J}}\|\tilde{J} - \tilde{J}_{n-1}\|_{\mathrm{F}} = \operatorname{argmin}_{\tilde{J}}\sqrt{\sum_{m,l=0}^{N-1}\left|[\tilde{J}]_{m,l} - [\tilde{J}_{n-1}]_{m,l}\right|^2}. \tag{3.176}$$

This solution has the closed form of a rank-one update with respect to the matrix of the previous iteration, namely [44][8]

$$\tilde{J}_n = \tilde{J}_{n-1} + \frac{\left(\Delta\tilde{G}^{(n)} - \tilde{J}_{n-1}\Delta z^{(n)}\right)\Delta z^{(n)\mathrm{H}}}{\|\Delta z^{(n)}\|^2}, \quad n = 2, 3, 4, \ldots \tag{3.177}$$

with

$$\Delta\tilde{G}^{(n)} = \tilde{G}(z^{(n)}) - \tilde{G}(z^{(n-1)}) \quad \text{and} \tag{3.178}$$

$$\Delta z^{(n)} = z^{(n)} - z^{(n-1)}. \tag{3.179}$$

By application of the Sherman-Morrison formula [29], we also have a rank-one update formula directly for the inverse:

$$\tilde{J}_n^{-1} = \tilde{J}_{n-1}^{-1} + \frac{\left(\Delta z^{(n)} - \tilde{J}_{n-1}^{-1}\Delta\tilde{G}^{(n)}\right)\Delta z^{(n)\mathrm{H}}\tilde{J}_{n-1}^{-1}}{\Delta z^{(n)\mathrm{H}}\tilde{J}_{n-1}^{-1}\Delta\tilde{G}^{(n)}}, \quad n = 2, 3, 4, \ldots \tag{3.180}$$

Obviously, these rank-one update formulas need an initial matrix, $\tilde{J}_1$, which is the original non-approximated full-rank Jacobian matrix (whose computation is expensive),

$$\tilde{J}_1 = \tilde{G}'(z^{(1)}). \tag{3.181}$$

Practical usage shows that we can gain up to approximately two times on computation time if we employ the Broyden's method when the computation of the Jacobian matrix is expensive. We cannot gain much more with this process in comparison to the Newton's method because, although each iteration is faster to execute, the convergence of Newton's method is quadratic and the convergence of Broyden's method, although supra-linear, is slower [44] (more iterations are

[8] In [44], the solution of a *real* non-linear equation system is considered, but it is easy to see that the following equation for G is equivalent to the one of the book for the real function $\tilde{F}$.

executed for the same precision). Note additionally that, as the full compressor function $\tilde{F}(x)$ is a real-valued real vector function, the Jacobian matrix of $\tilde{G}$ satisfies

$$D_N^{\mathrm{H}} \tilde{G}'(z) D_N \in \mathbb{R}^{N \times N}, \tag{3.182}$$

or equivalently, if a test vector $v \in \mathbb{C}^N$ is hermitian symmetric then $\tilde{G}'(z)v$ is hermitian symmetric as well. It is easy to prove that the same happens with $\tilde{G}'(z)^{-1}$, so that, using additionally the property that the dot-product of two hermitian symmetric vectors is real (prove!), the denominator of (3.180) is real for $n = 2$. Using the same property of $\tilde{G}'(z)$ and $\tilde{G}'(z)^{-1}$, we can also conclude that, for $n = 2$, both vectors which form the one-rank update in equations (3.177) and (3.180) (i.e., the vectors a and b which produce the update ab^{H}) are hermitian symmetric. The final consequence is that, for all $n = 2, 3, 4, \ldots$, the one-rank update vectors are hermitian symmetric and $D_N^{\mathrm{H}} \tilde{J}_n D_H$ is a real matrix.

3.5.3 Memory Optimization of the Expander

The developed expander uses a quasi-Newton method of the form (3.174) to fine-tune the initial estimate $z^{(1)}$. Unfortunately, this method has the inconvenient that the approximation for the Jacobian matrix $\tilde{J}_n$ takes a space of $O(N^2)$ in the memory of the computer that executes it. When the vector dimension that we are working with is large (typically we are working with value of N around 1024, but for theoretical simulations this number can go up to $N = 65536$), the memory used just to store that matrix is unbearably high. If, for example, we want to use $N = 65536$ just for the sake of theoretical simulations of the asymptotical behavior of the compressor and if one component of the matrix takes 32 bit of memory space, we will have to possess at least 16 GB RAM (Random Access Memory)! There is a solution to this problem, which is to avoid calculating $\tilde{J}_n$ explicitly, taking advantage of the Jacobian matrix of the compressor in the form of Equation (3.138).

Indeed, Equation (3.138) expresses the Jacobian Matrix as a low-rank (rank P) update of a cross-diagonal matrix (see Section 3.4.2 for an explanation on cross-diagonal matrices). We can exploit this form in the inversion of the matrix using the Woodbury matrix identity [27], here transcribed in a specialized form as

$$(V + XY^{\mathrm{H}})^{-1} = V^{-1} - V^{-1}X\left(\mathrm{I} + Y^{\mathrm{H}}V^{-1}X\right)^{-1}Y^{\mathrm{H}}V^{-1}, \tag{3.183}$$

for generic matrices $V \in \mathbb{C}^{N \times N}$ and X,$Y \in \mathbb{C}^{N \times P}$, with N and P arbitrary positive integers. Applying this identity to (3.138) delivers

$$\tilde{G}'_M(z)^{-1} = \frac{1}{\sqrt{N}} V(z)^{-1}\left(\mathrm{I} - A(z)C(z)^{-1}H(z)^{\mathrm{H}}V(z)^{-1}\right), \tag{3.184}$$

where

$$C(z) = \mathrm{I} + H(z)^{\mathrm{H}} V(z)^{-1} A(z). \tag{3.185}$$

Additionally, as proven in Appendix A (Section A.2), the inverse of a cross-diagonal matrix is cross-diagonal. Applying the result of the inverse in equations (A.15) to (A.17) to $V(z)$ in Equation (3.135), taking into account the hermitian symmetry of v_f and v_b and the fact that $v_f \in \mathbb{R}^N$, we get

$$V(z)^{-1} = \Lambda_{u_f(z)} + \Lambda_{u_b(z)} D_N^2 \tag{3.186}$$

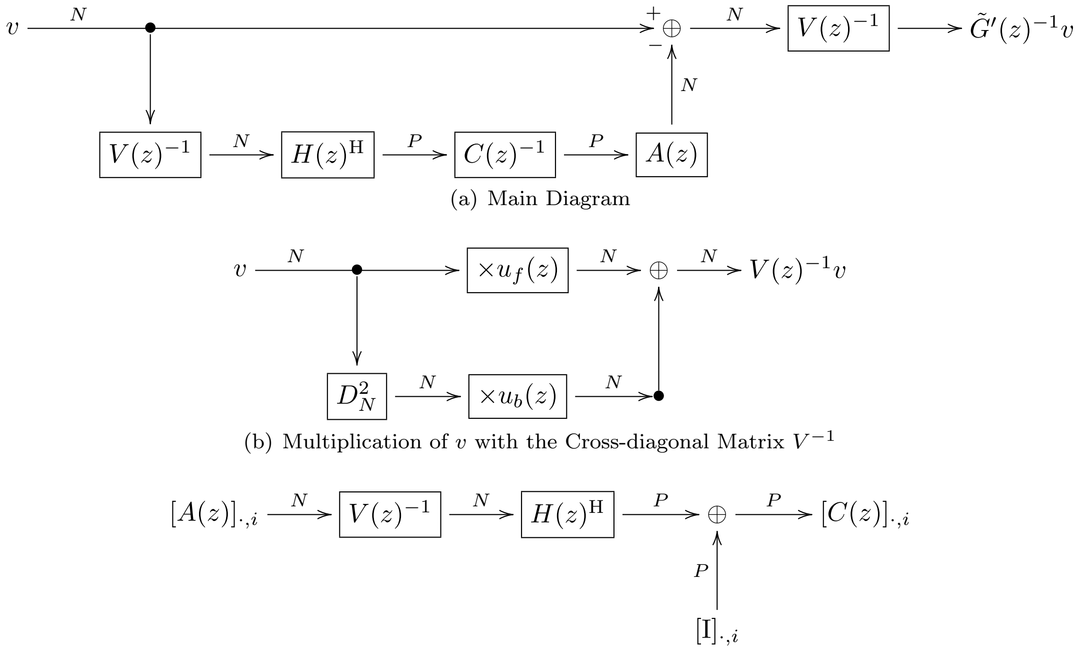
with

$$u_f(z,m) = \frac{1}{v_f(z,m)^2 - |v_b(z,m)|^2}\, v_f(z,m) \quad \text{and} \tag{3.187}$$

$$u_b(z,m) = -\frac{1}{v_f(z,m)^2 - |v_b(z,m)|^2}\, v_b(z,m). \tag{3.188}$$

Note that V^{-1} also has the property that u_f and u_b are hermitian symmetric (and u_f is even real), so that a left multiplication of a hermitian symmetric vector by V^{-1} produces a hermitian symmetric vector as well. Additionally, due to the hermitian symmetry of the columns of A and H, (3.185) is a real (P-by-P) matrix. Finally, note that $u_b(z,0) = u_b(z,N/2)$ due to the lower branch of Equation (3.137).

We can thus implement any multiplication of the matrix $\tilde{G}'_M(z)^{-1}$ with any vector v with the block diagrams of Figure 3.2. As you can see from the main diagram, Figure 3.2(a), we first multiply the input vector v with the cross-diagonal matrix V^{-1}. Although this is a large N-by-N matrix, we do not need to calculate it explicitly, as it is sparse. Indeed, as shown in Equation (3.186) and in the block diagram of Figure 3.2(b), for performing this multiplication we only need to swap the input vector, i.e., do the operation $[v(0), v(1), v(2), \dots, v(N-1)]^{\mathrm{T}} \rightarrow [v(0), v(N-1), v(N-2), \dots, v(1)]^{\mathrm{T}}$ (this is the left multiplication by D_N^2; cf. matrix (3.60)), then multiply the result point-wise by u_b, and sum it to the point-wise multiplication of the original vector by u_f. The second step in the main diagram is the left multiplication of the result of $V^{-1}v$ by the P-by-N matrix H^{H}, producing a vector of size P. Note that usually $P \ll N$ and that this resulting vector is real in the case of interest, where v is hermitian symmetric. We then multiply this last result by a P-by-P matrix C^{-1}, and finally by the N-by-P matrix A, returning to the original dimension N. That result (which is hermitian symmetric in the case of interest) is subtracted from the original input vector and the output is multiplied again with V^{-1} (again with the diagram of Figure 3.2(b)), being thus (3.184) implemented without recurring to any N-by-N matrix. The maximum matrix size that we need is N-by-P, and taking the same example as above ($N = 65536$) with $P = 60$, we only need 15 MB $\times$ 2 to store the matrices A and H now (the other elements have a negligible memory consumption with respect to this one). Note that for the calculation of C, we do not need to use any N-by-N either; we only need to run the diagram of Figure 3.2(c) for every column of C (P times), and this scheme runs operations which are similar to the ones described above. The inversion of C for the execution of Figure 3.2(a) is either done explicitly (the matrix is small; its size is namely P-by-P) once, or we can solve the linear equation system $C(z)\,v = r$ every time we want to do an operation of the type $v = C(z)^{-1}r$, $v,r \in \mathbb{R}^P$.



(a) Main Diagram

(b) Multiplication of v with the Cross-diagonal Matrix V^{-1}

(c) Building the column i of C, $i = 0, 1, \ldots, P-1$

Figure 3.2: Block Diagrams for the execution of operation $\tilde{G}'(z)^{-1}v$ for some (hermitian symmetric) vector v. The labels of the arrows denote the size of the vector output on the previous block, × denotes point-wise multiplication and • are simple junction points.

To apply the Broyden's method using the cheaper way to compute the matrix multiplication $\tilde{G}'(z)^{-1}v$ described in the last paragraph, it is easy to see that if we define

$$a^{(n)} = \frac{\Delta z^{(n)} - \tilde{J}_{n-1}^{-1}\Delta\tilde{G}^{(n)}}{\Delta z^{(n)\mathrm{H}}\tilde{J}_{n-1}^{-1}\Delta\tilde{G}^{(n)}} \quad \text{and} \tag{3.189}$$

$$b^{(n)} = \tilde{J}_{n-1}^{-\mathrm{H}}\Delta z^{(n)} \tag{3.190}$$

then the approximation of the Jacobian matrix in Broyden's method, (3.180) and (3.181), can be written non-recursively as

$$\tilde{J}_n^{-1} = \tilde{G}'_M(z^{(1)})^{-1} + \sum_{k=2}^{n} a^{(k)}\left[b^{(k)}\right]^{\mathrm{H}}, \quad n = 1, 2, 3, \ldots \tag{3.191}$$

with $\sum_{k=2}^{1} \equiv 0$. If we keep all $a^{(k)}, b^{(k)}$, $k = 2, 3, \ldots n$ on memory, we can thus calculate any

$\tilde{J}_n^{-1}v$ (for hermitian symmetric v, cf. Equation (3.174)) by implementing

$$\tilde{J}_n^{-1}v = \tilde{G}'_M(z^{(1)})^{-1}v + \sum_{k=2}^{n} a^{(k)} \left(\left[b^{(k)} \right]^{\mathrm{H}} v \right) \tag{3.192}$$

with the methodology described above for $\tilde{G}'(z^{(1)})^{-1}v$, i.e., we accumulate, for $k = 2, 3, \ldots, n$, the dot-product of v with $[b^{(k)}]^{\mathrm{H}}$ (a real scalar) times the vector $a^{(k)}$ (a hermitian symmetric vector), and sum the result with the inverse of the derivative of $\tilde{G}_M$ at $z^{(1)}$, applied to v using the diagram of Figure 3.2. To build the new $a^{(n)}$, Equation (3.189), when passing from iteration $n-1$ to the iteration n, we can use (3.192) for $\tilde{J}_{n-1}^{-1}$ with $v = \Delta\tilde{G}^{(n)}$. For $b^{(n)}$, we do a similar procedure with $\tilde{J}_n^{-\mathrm{H}}$. Applying the hermitian conjugate operator to (3.191) and (3.184) we get

$$\tilde{J}_n^{-\mathrm{H}} = \tilde{G}'(z^{(1)})^{-\mathrm{H}} + \sum_{k=2}^{n} b^{(k)} \left[a^{(k)} \right]^{\mathrm{H}}, \quad n = 1, 2, 3, \ldots \tag{3.193}$$

$$\tilde{J}_n^{-\mathrm{H}}v = \tilde{G}'(z^{(1)})^{-\mathrm{H}}v + \sum_{k=2}^{n} b^{(k)} \left(\left[a^{(k)} \right]^{\mathrm{H}} v \right) \tag{3.194}$$

and

$$\tilde{G}'_M(z)^{-\mathrm{H}} = \frac{1}{\sqrt{N}} \left(\mathrm{I} - V(z)^{-\mathrm{H}} H(z) C(z)^{-\mathrm{H}} A(z)^{\mathrm{H}} \right) V(z)^{-\mathrm{H}} \tag{3.195}$$

$$= \frac{1}{\sqrt{N}} \left(\mathrm{I} - V(z)^{-1} H(z) C(z)^{-\mathrm{T}} A(z)^{\mathrm{H}} \right) V(z)^{-1} \tag{3.196}$$

$$= \frac{1}{\sqrt{N}} V(z)^{-1} \left(\mathrm{I} - H(z) C(z)^{-\mathrm{T}} A(z)^{\mathrm{H}} V(z)^{-1} \right), \tag{3.197}$$

where in (3.196) we used the fact that u_f is real and that u_b is hermitian symmetric to state $V^{-\mathrm{H}} = V^{-1}$ and also the fact that C is a real matrix. We can thus calculate $b^{(n)}$ using (3.194) for $\tilde{J}_{n-1}^{-\mathrm{H}}$ and (3.197) with the operation order described in the text corresponding to Figure 3.2. These equations are very similar to the ones of $\tilde{J}_{n-1}^{-1}$, with the difference that a gets switched by b, A by H and vice-versa, and we use $C^{-\mathrm{T}}$ instead of C^{-1}. The diagram that we get for this operation is the same as the diagram of Figure 3.2, up to these differences.

Chapter 4

Simulating the Suboptimal Compressor

A suboptimal companding scheme was developed and analyzed in Chapter 3, and it was proven that it behaves optimally when the vector dimension N goes to infinity. In this chapter, we simulate the companding scheme, showing simulation figures and tables that corroborate the theoretical results. The main simulation compares rate-distortion figures associated to the distortion measure with the actual rate-distortion performance obtained by the compander. The simulations are done for the case of high resolution, that is, when the size of the basic element of the quantization lattice is much smaller than the standard deviation of the source ($D \to 0$). This approximation is valid when we are coding audio at high fidelity (at about 32 kbps per channel or higher), which is the case of interest for this thesis. For doing such simulations, we first calculate the rate-distortion function for the distortion measure (and for a certain source) we are concerned with (see Section 2.3), and afterwards, we also calculate the rate-loss (2.39). Following these calculations, we show and discuss the executed simulations.

It is important to note that, although, in most of the cases, the correspondent results are not shown in this thesis, every step in Chapter 3 (and specifically every approximation) was confirmed either by calculating derivatives (e.g., the sensitivity matrix or the Jacobian matrix of the compressor) numerically for a white noise test signal (i.e., using the difference quotient) and overlaying the numerical results with the calculated ones, or by confirming that numerical optimizations delivered exactly the same outputs as the non-optimized versions, up to numerical roundoff errors.

4.1 The Rate Distortion Function at High Resolution

In this section, we calculate the rate-distortion function $R(D)$ explicitly for the distortion measure (2.42) and examine its behavior for $N \to \infty$.

4.1.1 Calculation

We will now calculate the rate-distortion function (2.6) for the distortion measure (2.42). It is easy to see that this distortion measure is locally quadratic (conditions (2.1) and (2.3) apply) and that the remaining more technical conditions described in [37] also apply for non-pathological sources due to the invertibility of the sensitivity matrix. We are thus allowed to use the result for the rate-distortion function for a locally quadratic distortion measure (2.10) (see Section 2.1).

To calculate the last term of the Equation (2.10), we will use the case $L \neq N$ as reference but, as the equations defining $M(x)$ in the case $L = N$ (equations (3.9) and (3.17)) are exactly the same as the ones of the approximation of the Toeplitz matrix $M_t(x)$ by the circulant matrix $\bar{M}_c(x)$ in the case $L = N$ (equations (3.54) and (3.37)), we can use the following results for both $L \neq N$ and $L = N$ ($\bar{M}_c(x)$ and $M_t(x)$ degenerate both in M_c and $\bar{a}$ degenerates in a in this last case). We get, calculating thus the determinant of $M(x)$ for $L \neq N$,

$$\det M(x) = \det[\Lambda_w M_t(x) \Lambda_w] = \left[\prod_{n=0}^{N-1} w(n)\right]^2 \det M_t(x) \tag{4.1}$$

$$\approx \left[\prod_{n=0}^{N-1} w(n)\right]^2 \prod_{f=0}^{N-1} [N\hat{\bar{a}}^2(x,f)], \tag{4.2}$$

where (4.1) is the exact value of the determinant for $L \neq N$ and (4.2) is the approximated value. Both equations are exact for $L = N$. Extracting the binary logarithm of the determinant gives

$$\log_2 \det M(x) = 2\sum_{n=0}^{N-1} \log_2 w(n) + \log_2 \det M_t(x) \tag{4.3}$$

$$\approx 2\sum_{n=0}^{N-1} \log_2 w(n) + N \log_2 N + \sum_{f=0}^{N-1} \log_2[\hat{\bar{a}}^2(x,f)], \tag{4.4}$$

so that the last term of the rate-distortion function (2.10) is

$$\frac{1}{2N}\mathrm{E}[\log_2 \det M(X)] = \frac{1}{N}\sum_{n=0}^{N-1} \log_2 w(n) + \frac{1}{2N}\mathrm{E}\log_2 \det M_t(X) \tag{4.5}$$

$$\approx \frac{1}{N}\sum_{n=0}^{N-1} \log_2 w(n) + \frac{\log_2 N}{2} + \frac{1}{2N}\sum_{f=0}^{N-1} \mathrm{E}\log_2[\hat{\bar{a}}^2(X,f)]. \tag{4.6}$$

Summing up equations (2.10) and (4.6), the rate-distortion function for the distortion measure

(2.42) is given, at high resolution, by

$$R(D) \approx h(X) - \frac{1}{2}\log_2(2\pi\mathrm{e}D) + \frac{1}{N}\sum_{n=0}^{N-1}\log_2 w(n) + \frac{1}{2N}\mathrm{E}\log_2\det M_t(X) \tag{4.7}$$

$$\approx h(X) - \frac{1}{2}\log_2(2\pi\mathrm{e}D) + \frac{1}{N}\sum_{n=0}^{N-1}\log_2 w(n) + \frac{\log_2 N}{2} + \frac{1}{2N}\sum_{f=0}^{N-1}\mathrm{E}\log_2[\hat{\bar{a}}^2(X,f)] \tag{4.8}$$

and the distortion-rate functions, which can be obtained using Equation (2.11), have analog expressions.

Obviously, in practice, when implementing Equation (4.7) or (4.8) we have to substitute the expectation operator by a statistical average, with increasingly accurate results as the number of (N-sized) x sample vectors increases (assuming that the x vectors are independent).

It should be noted that for the case $L \neq N$, the approximation $M(X) \approx \bar{M}(X)$ produces asymptotically an exact result in (4.6), since from (3.54), (3.30) and (3.56) with $\Psi(v) = \log_2 v$, we have

$$\lim_{N\to\infty}\frac{1}{2N}\mathrm{E}[\log_2\det\bar{M}(X)] = \lim_{N\to\infty}\frac{1}{2N}\mathrm{E}[\log_2\det\bar{M}_c(X)] + \lim_{N\to\infty}\frac{1}{N}\log_2\det\Lambda_w \tag{4.9}$$

$$= \mathrm{E}\left[\lim_{N\to\infty}\frac{1}{2N}\sum_{f=0}^{N-1}\log_2 e_c(X,f)\right] + \lim_{N\to\infty}\frac{1}{N}\log_2\det\Lambda_w \tag{4.10}$$

$$= \mathrm{E}\left[\lim_{N\to\infty}\frac{1}{2N}\sum_{f=0}^{N-1}\log_2 e_t(X,f)\right] + \lim_{N\to\infty}\frac{1}{N}\log_2\det\Lambda_w \tag{4.11}$$

$$= \lim_{N\to\infty}\frac{1}{2N}\mathrm{E}[\log_2\det M_t(X)] + \lim_{N\to\infty}\frac{1}{N}\log_2\det\Lambda_w \tag{4.12}$$

$$= \lim_{N\to\infty}\frac{1}{2N}\mathrm{E}[\log_2\det M(X)]. \tag{4.13}$$

4.1.2 Asymptotic Expression ($N \to \infty$) and Behavior

Although numerically irrelevant, theoretically it is of interest to see that for N much larger than the support of the autocorrelation of a and than the support of $\hat{w}$, we can approximate sums by integrals in Equation (4.6), since then the variations of $\hat{\bar{a}}^2$ and w take place on a much larger time- and frequency-scale than $1/N$, respectively. In fact, denoting by $\hat{\bar{a}}^2(x,\nu)$ the continuous inverse of the masking threshold, given by the limit as $L \to \infty$ of Equation (2.44) (with re-calibration of c_2 with changing L) at the linear frequency value of $\nu = f/L \cdot f_s$, and denoting also by $w(t)$ the continuous window function defined in the interval $[0, 1]$, obtained by calculating a window of size

$N \to \infty$ and doing the substitution $t = n/N$, we get as an approximation

$$\frac{1}{2N}\mathrm{E}[\log_2 \det M(X)] = \sum_{n=0}^{N-1} \log_2[w(n)]\,\frac{1}{N} + \frac{\log_2 N}{2} + \frac{1}{2f_s}\sum_{f=0}^{N-1} \mathrm{E}\log_2[\hat{\bar{a}}^2(X,f)]\,\frac{f_s}{N} \tag{4.14}$$

$$\approx \int_0^1 \log_2 w(t)\,\mathrm{d}t + \frac{\log_2 N}{2} + \frac{1}{2f_s}\int_0^{f_s} \mathrm{E}\log_2[\hat{\bar{a}}^2(X,\nu)]\,\mathrm{d}\nu \tag{4.15}$$

$$= \int_0^1 \log_2 w(t)\,\mathrm{d}t + \frac{\log_2 N}{2} + \frac{1}{f_s}\int_0^{\frac{f_s}{2}} \mathrm{E}\log_2[\hat{\bar{a}}^2(X,\nu)]\,\mathrm{d}\nu, \tag{4.16}$$

where in (4.16) we used the even symmetry of $\hat{\bar{a}}^2(x,\nu)$ around $\nu = f_s/2$.

The variant of the inverse of the masking threshold $\hat{a}'^2(x,\cdot)$ exposed in Subsection 2.3.2 is useful to study the behavior of $R(D)$ when $N \to \infty$, since $\hat{a}'^2(x,\cdot)/N$ is asymptotically independent of N. Indeed, using equations (2.56) and (3.49), we get the succession of equalities

$$\hat{\bar{a}}^2(x,f) = \frac{L}{N}\hat{a}^2\left(x, \frac{L}{N}f\right) = \frac{1}{N^2}\hat{a}'^2\left(x, \frac{L}{N}f\right). \tag{4.17}$$

We can thus rewrite Equation (4.6) in terms of $\hat{a}'^2(x,\cdot)/N$, getting

$$\frac{1}{2N}\mathrm{E}[\log_2 \det M(X)] = \frac{1}{N}\sum_{n=0}^{N-1} \log_2 w(n) + \frac{1}{2N}\sum_{f=0}^{N-1} \mathrm{E}\log_2\left[\frac{\hat{a}'^2\left(X, \frac{L}{N}f\right)}{N}\right]. \tag{4.18}$$

Finally, following the same steps as in (4.14) through (4.16), we have

$$\frac{1}{2N}\mathrm{E}[\log_2 \det M(X)] \approx \int_0^1 \log_2 w(t)\,\mathrm{d}t + \frac{1}{f_s}\int_0^{\frac{f_s}{2}} \mathrm{E}\log_2\left[\frac{\hat{a}'^2(X,\nu)}{N}\right]\mathrm{d}\nu, \tag{4.19}$$

where $\hat{a}'^2(x,\nu) = NL\,\hat{a}^2(x,\nu)$, with the rate-distortion function at high resolution given by

$$R(D) \approx h(X) - \frac{1}{2}\log_2(2\pi \mathrm{e} D) + \int_0^1 \log_2 w(t)\,\mathrm{d}t + \frac{1}{f_s}\int_0^{\frac{f_s}{2}} \mathrm{E}\log_2\left[\frac{\hat{a}'^2(X,\nu)}{N}\right]\mathrm{d}\nu. \tag{4.20}$$

Equation (4.20) converges when $N \to \infty$. As previously explained, the last term involving $\hat{a}'^2(x,\nu)/N$ has its dependence canceled out asymptotically, and similar arguments apply to the distortion per dimension D in the first term (cf. Subsection 2.3.2). This means that, for N not too small, so that the autocorrelation of the inverse of the masking threshold has a support much smaller than N and that the power estimates of $x * h_i$ and $(y - x) * h_i$ have almost converged, the rate distortion function is practically independent of N, that is, if we want to know the minimal rate associated with the distortion per dimension D, we know that that rate is the same[1] independently of N.

[1] Naturally, this only applies for $N/f_s < 300$ ms, since otherwise $\hat{a}'^2$ is not proportional to N.

4.2 Rate-loss of the Suboptimal Compander

For doing the main simulation of this thesis, besides needing a basis for comparison given by the best achievable (the rate-distortion function), we also need to know how the actual scheme performs in a rate-distortion sense, or equivalently, how much we loose with respect to the rate-distortion curve. That loss is characterized by Equation (2.39), and will be calculated in this section. We will first calculate the rate-loss for the case that we do not use any companding scheme, using the identity function as the compressor and expander ($F \equiv \mathrm{I}$, $F(x) = x$), and then recalculate it for the compressor of Chapter 3.

4.2.1 Identity Compander

For the identity compander $F(x) = F(x)^{-1} = x$, the Jacobian matrix of F and $\tilde{M}$ of Equation (2.38) are both the identity matrix and the rate-loss (at high resolution) degenerates in

$$H(Q_{D,\mathrm{I}}) - H(Q_{D,F}) \approx -\frac{1}{2N}\mathrm{E}\log_2 \det M(X) + \frac{1}{2}\log_2\left[\mathrm{E}\ \mathrm{tr}\left(\frac{M(X)}{N}\right)\right]. \tag{4.21}$$

The first term on the right hand side was already calculated in Section 4.1. For the second term, due to the linearity of the trace operator, we just need to calculate tr M and divide the result by N afterwards. We have then from equations (3.30), (3.54), and (3.37)

$$\mathrm{tr}\ M(x) = \mathrm{tr}\left[\Lambda_w M_t(x)\Lambda_w\right] = \mathrm{tr}\left[\Lambda_w^2 M_t(x)\right] = \left[\sum_{n=0}^{N-1} w^2(n)\right]\left[\frac{1}{L}\sum_{f=0}^{L-1} L\hat{a}^2(x,f)\right] \tag{4.22}$$

$$\approx \mathrm{tr}\left[\Lambda_w^2 D_N^{\mathrm{H}}\Lambda_{N\hat{a}^2} D_N\right] = \left[\sum_{n=0}^{N-1} w^2(n)\right]\left[\frac{1}{N}\sum_{f=0}^{N-1} N\hat{a}^2(x,f)\right], \tag{4.23}$$

where we used the fact that M_t and $\bar{M}_c$ are Toeplitz matrices, with equal elements along the diagonal, given by the inverse DFT of the eigenvalues of M_c at the time-index $n = 0$. We have thus, using (4.5),

$$\begin{aligned} H(Q_{D,\mathrm{I}}) - H(Q_{D,F}) \approx &-\frac{1}{N}\sum_{n=0}^{N-1}\log_2 w(n) - \frac{1}{2N}\mathrm{E}\log_2\det M_t(X) \\ &+\frac{1}{2}\log_2\left[\frac{1}{N}\sum_{n=0}^{N-1} w^2(n)\right] + \frac{1}{2}\log_2\mathrm{E}\left[\frac{1}{L}\sum_{f=0}^{L-1} L\hat{a}^2(X,f)\right] \qquad (4.24) \\ \approx &-\frac{1}{N}\sum_{n=0}^{N-1}\log_2 w(n) - \frac{1}{2N}\sum_{f=0}^{N-1}\mathrm{E}\log_2[N\hat{a}^2(X,f)] \\ &+\frac{1}{2}\log_2\left[\frac{1}{N}\sum_{n=0}^{N-1} w^2(n)\right] + \frac{1}{2}\log_2\mathrm{E}\left[\frac{1}{N}\sum_{f=0}^{N-1} N\hat{a}^2(X,f)\right], \qquad (4.25) \end{aligned}$$

where, for $L \neq N$, the upper equation is valid exactly and the lower equation refers to the approximation of M_t by $\bar{M}_c$. For $L = N$, both equations deliver the exact results ($\bar{M}_c$ and M_t degenerate both in M_c and $\bar{a}$ degenerates in a). To retrieve this loss in terms of distortion, we can simply use (2.41) for this particular rate-loss of Equation (4.25).

The behavior of the rate-loss when $N \to \infty$ follows similar steps as the ones in Subsection 4.1.2: we use the dimension independent $\hat{a}'^2(x,f)/N$ instead of $\hat{\bar{a}}^2(x,f)$ (cf. Subsection 2.3.2) and approximate sums of discrete functions by integrals of their continuous counterparts. The result is

$$\begin{aligned} H(Q_{D,\mathrm{I}}) - H(Q_{D,F}) \approx & -\int_0^1 \log_2 w(t)\,\mathrm{d}t - \frac{1}{f_s}\int_0^{\frac{f_s}{2}} \mathrm{E}\log_2\left[\frac{\hat{a}'^2(X,\nu)}{N}\right]\mathrm{d}\nu \\ & + \frac{1}{2}\log_2\int_0^1 w^2(t)\,\mathrm{d}t + \frac{1}{2}\log_2\left[\frac{2}{f_s}\int_0^{\frac{f_s}{2}} \mathrm{E}\left\{\frac{\hat{a}'^2(X,\nu)}{N}\right\}\mathrm{d}\nu\right]. \end{aligned} \tag{4.26}$$

Note that from Jensen's inequality, the following chains are valid:

$$\int_0^1 \log_2 w(t)\,\mathrm{d}t = \frac{1}{2}\int_0^1 \log_2 w^2(t)\,\mathrm{d}t \tag{4.27}$$

$$\leq \frac{1}{2}\log_2\left[\int_0^1 w^2(t)\,\mathrm{d}t\right] \tag{4.28}$$

and

$$\frac{1}{f_s}\int_0^{\frac{f_s}{2}} \mathrm{E}\left\{\log_2\left[\frac{\hat{a}'^2(X,\nu)}{N}\right]\right\}\mathrm{d}\nu = \frac{1}{2f_s}\int_0^{f_s} \mathrm{E}\left\{\log_2\left[\frac{\hat{a}'^2(X,\nu)}{N}\right]\right\}\mathrm{d}\nu \tag{4.29}$$

$$\leq \frac{1}{2f_s}\int_0^{f_s} \log_2 \mathrm{E}\left\{\frac{\hat{a}'^2(X,\nu)}{N}\right\}\mathrm{d}\nu \tag{4.30}$$

$$\leq \frac{1}{2}\log_2\left[\frac{1}{f_s}\int_0^{f_s} \mathrm{E}\left\{\frac{\hat{a}'^2(X,\nu)}{N}\right\}\mathrm{d}\nu\right] \tag{4.31}$$

$$= \frac{1}{2}\log_2\left[\frac{2}{f_s}\int_0^{\frac{f_s}{2}} \mathrm{E}\left\{\frac{\hat{a}'^2(X,\nu)}{N}\right\}\mathrm{d}\nu\right]. \tag{4.32}$$

These chains confirm that the rate-loss (4.26) is greater or equal than 0 and make explicit that we have to satisfy the equalities in it to achieve optimality. Nevertheless, $\log_2$ is a strictly concave function and, in general, the distribution of $\hat{a}'^2(X,\nu)$ is not degenerate (its probability density function is not a Dirac delta distribution). Consequently, equality is, in general, not achieved in the series of inequalities (4.27) to (4.32) so that we cannot achieve optimality using the identity compressor, even with $N \to \infty$.

4.2.2 Compander of Chapter 3

We will also use the decomposition (3.138) of the Jacobian matrix of the compressor here to avoid calculating large N-by-N matrices explicitly. From Equation (2.39), we see that the new terms that we have to calculate are the expected values of $[\log_2 \det \tilde{M}]/(2N)$ and of $\operatorname{tr}(\tilde{M}^{-1}M)/N$. For the first expression, algebraic manipulation using equations (2.38) and (3.65) delivers

$$\frac{1}{2N}\log_2 \det \tilde{M}(x) = \frac{1}{2N}\log_2 \det[\tilde{F}'(x)^{\mathrm{T}}\tilde{F}'(x)] = \frac{1}{N}\log_2 |\det \tilde{F}'(x)| \tag{4.33}$$

$$= \frac{1}{N}\log_2 |\det[D_N^H\,\tilde{G}'(z)\,D_N\Lambda_w]| = \frac{1}{N}\sum_{n=0}^{N-1}\log_2 w(n) + \frac{1}{N}\log_2 |\det \tilde{G}'(z)| \tag{4.34}$$

with z equal to the one of Equation (3.20). Applying the matrix determinant lemma to (3.138), we obtain

$$\frac{1}{N}\log_2 |\det \tilde{G}'(z)| = \frac{1}{N}\log_2 |\det[\sqrt{N}(V(z) + A(z)H(z)^{\mathrm{H}})]| \tag{4.35}$$

$$= \frac{\log_2 N}{2} + \frac{1}{N}\log_2 |\det V(z)| + \frac{1}{N}\log_2 |\det[\mathrm{I} + H(z)^{\mathrm{H}}V(z)^{-1}A(z)]| \tag{4.36}$$

$$= \frac{\log_2 N}{2} + \frac{1}{2N}\sum_{m=0}^{N-1}\log_2 \left|v_f(z,m)^2 - |v_b(z,m)|^2\right| + \frac{1}{N}\log_2 |\det C(z)|, \tag{4.37}$$

where we used (3.185), (3.135), (A.14), and the symmetry properties of v_f and v_b.

For $\operatorname{tr}(\tilde{M}^{-1}M)$, we can use simple properties of linear algebra, namely the commutativity of the trace of a product of two square matrices, to get

$$\operatorname{tr}[\tilde{M}(x)^{-1}M(x)] = \operatorname{tr}\left\{[\tilde{F}'(x)^{\mathrm{T}}\tilde{F}'(x)]^{-1}M(x)\right\} \tag{4.38}$$

$$= \operatorname{tr}\left\{[\Lambda_w D_N^{\mathrm{H}}\tilde{G}'(z)^{\mathrm{H}}\tilde{G}'(z)D_N\Lambda_w]^{-1}\Lambda_w M_t(x)\Lambda_w\right\} \tag{4.39}$$

$$= \operatorname{tr}\left\{\left(\Lambda_w^{-1}D_N^{\mathrm{H}}\right)\left(\tilde{G}'(z)^{-1}\tilde{G}'(z)^{-\mathrm{H}}D_N M_t(x)\Lambda_w\right)\right\} \tag{4.40}$$

$$= \operatorname{tr}\left\{\tilde{G}'(z)^{-1}\tilde{G}'(z)^{-\mathrm{H}}D_N M_t(x)D_N^{\mathrm{H}}\right\} = \operatorname{tr}\left\{\tilde{G}'(z)^{-1}\tilde{G}'(z)^{-\mathrm{H}}\hat{M}_t(x)\right\} \tag{4.41}$$

with

$$\hat{M}_t(x) \overset{\mathrm{def}}{=} D_N M_t(x) D_N^{\mathrm{H}} \tag{4.42}$$

equal to the diagonal matrix $\Lambda_{N\hat{a}^2}$ for $L = N$, $\Lambda_{N\hat{\bar{a}}^2}$ when approximating the Toeplitz matrix M_t by the circulant one (the matrix $\bar{M}_c$) for $L \neq N$, and if we do not want to do that approximation for $L \neq N$, a certain non-diagonal matrix with the given expression. We will use the Taylor expanded compressor $\tilde{G}_M$ from now on, replacing $\tilde{G}'$ by $\tilde{G}'_M$ in the previous expressions. From equations (3.184) and (3.196), we can continue working out this last result:

$$\begin{aligned}\operatorname{tr}[\tilde{M}(x)^{-1}M(x)] = \frac{1}{N}\operatorname{tr}\{ &\left(\mathrm{I} - A(z)C(z)^{-1}H(z)^{\mathrm{H}}V(z)^{-1}\right) \cdot \\ &\cdot \left(\mathrm{I} - V(z)^{-1}H(z)C(z)^{-\mathrm{T}}A(z)^{\mathrm{H}}\right) M_u(z)\},\end{aligned} \tag{4.43}$$

where we define the unwindowed sensitivity matrix M_t left and right multiplied by V^{-1} by M_u:

$$M_u(z) \overset{\mathrm{def}}{=} V(z)^{-1}\hat{M}_t(x)V(z)^{-1}. \tag{4.44}$$

We have then

$$\begin{aligned}\operatorname{tr}[\tilde{M}(x)^{-1}M(x)] = &\frac{1}{N}\operatorname{tr}\{M_u(z)\} - \frac{1}{N}\operatorname{tr}\left\{C(z)^{-1}H(z)^{\mathrm{H}}V(z)^{-1}M_u(z)A(z)\right\} \\ &- \frac{1}{N}\operatorname{tr}\left\{C(z)^{-\mathrm{T}}A(z)^{\mathrm{H}}M_u(z)V(z)^{-1}H(z)\right\} \\ &+ \frac{1}{N}\operatorname{tr}\left\{C(z)^{-1}H(z)^{\mathrm{H}}V(z)^{-2}H(z)C(z)^{-\mathrm{T}}A(z)^{\mathrm{H}}M_u(z)A(z)\right\}.\end{aligned} \tag{4.45}$$

The first subtractive term on the right hand side of (4.45) is real: as the matrix $V^{-1}M_u$ is a product of matrices which produce a hermitian symmetric vector when left multiplying hermitian symmetric vectors (this is true for $\hat{M}_t$ due to the form (4.42) with a real M_t), that matrix itself has the same property. The columns of A are hermitian symmetric, so that the columns of $V^{-1}M_uA$ are hermitian symmetric as well and $H^{\mathrm{H}}V^{-1}M_uA$ is real, as a consequence of being a matrix composed by dot-products of hermitian symmetric vectors (H has hermitian symmetric columns as well). C also is real, proving that the result of the multiplication $C^{-1}H^{\mathrm{H}}V^{-1}M_uA$ is real, thus that this last matrix has a real trace. Furthermore, the two subtractive terms of (4.45) are equal:

$$\begin{aligned}\operatorname{tr}\left\{C(z)^{-\mathrm{T}}A(z)^{\mathrm{H}}M_u(z)V(z)^{-1}H(z)\right\} &= \operatorname{tr}\left\{\left[C(z)^{-\mathrm{T}}A(z)^{\mathrm{H}}M_u(z)V(z)^{-1}H(z)\right]^{\mathrm{H}}\right\} && (4.46)\\ &= \operatorname{tr}\left\{H(z)^{\mathrm{H}}V(z)^{-1}M_u(z)A(z)C(z)^{-1}\right\} && (4.47)\\ &= \operatorname{tr}\left\{C(z)^{-1}H(z)^{\mathrm{H}}V(z)^{-1}M_u(z)A(z)\right\}, && (4.48)\end{aligned}$$

where we used $V^{-\mathrm{H}} = V^{-1}$ and the symmetry of the real matrix M_t to state that

$$M_u(z)^{\mathrm{H}} = V(z)^{-1}D_N M_t(x)^{\mathrm{H}}D_N^{\mathrm{H}}V(z)^{-1} = V(z)^{-1}D_N M_t(x)D_N^{\mathrm{H}}V(z)^{-1} = M_u(z). \tag{4.49}$$

We can thus leave out the third term of Equation (4.45) if we double the second one:

$$\begin{aligned}\operatorname{tr}[\tilde{M}(x)^{-1}M(x)] = &\frac{1}{N}\operatorname{tr}\{M_u(z)\} - \frac{2}{N}\operatorname{tr}\left\{C(z)^{-1}\left[H(z)^{\mathrm{H}}\left(V(z)^{-1}M_u(z)\right)A(z)\right]\right\} \\ &+ \frac{1}{N}\operatorname{tr}\left\{C(z)^{-1}\left[H(z)^{\mathrm{H}}V(z)^{-2}H(z)\right]C(z)^{-\mathrm{T}}\left[A(z)^{\mathrm{H}}M_u(z)A(z)\right]\right\}.\end{aligned} \tag{4.50}$$

The parentheses in Equation (4.50) suggest the best order in which we should compute the trace of $\tilde{M}^{-1}M$ when implementing the rate-loss equations. All matrices on the level of the brackets $[\cdot]$ on the last two terms of the right hand side of (4.50) are P-by-P, so that the shown multiplication is computationally cheap to do. To compute those P-by-P matrices, we first have to left multiply an N-by-P matrix (A or H) by a matrix which will be shown to be cross-diagonal for the cases $L = N$ and $L \neq N$ when approximating M_t by the circulant matrix $\bar{M}_c$. As shown in Subsection 3.5.3 (see specifically Figure 3.2(b)), the left multiplication by a cross diagonal matrix is fast and we do not need to store the cross-diagonal matrix explicitly. In this case, we have to perform the multiplication using a scheme similar to the one of Figure 3.2(b) P times, for all columns of A or H. The only case where the matrix in question is not cross-diagonal is the case where we want to calculate the rate-loss for $L \neq N$ without using the approximation by a circulant matrix. As it is only worth to use that case for small N (since the approximation is asymptotically correct), there are no memory or computational complexity problems in handling N-by-N matrices, so that we can calculate all N-by-N matrices explicitly. After the computationally feasible left multiplication of A or H by an N-by-N matrix, we only have to multiply the result by a P-by-N matrix (H^{H} or A^{H}), which implies computing P^2 dot products of N-sized vectors, also a computationally feasible task. We thus conclude (if we prove the cross-diagonality of the N-by-N matrices in the cases where we use large N) that the execution of (4.50) is feasible even for large N, e.g., $N = 65536$.

On the right hand side of Equation (4.50), the only N-by-N matrix we have left on the outer level (the level of the brackets $[\cdot]$) is M_u, in the first term. It is also not necessary to calculate this matrix explicitly for large N, since it is further simplifiable by using the commutative property of the trace in Equation (4.44). We can easily obtain

$$\operatorname{tr}\{M_u(z)\} = \operatorname{tr}\left\{V(z)^{-2}\hat{M}_t(x)\right\} \tag{4.51}$$

and using the simple to deduce properties

$$D_N^2\Lambda_v = \Lambda_{D_N^2 v}D_N^2, \tag{4.52}$$

$$D_N^4 = \mathrm{I} \tag{4.53}$$

and the symmetry properties of u_f and u_b, we calculate V^{-2} explicitly as

$$V(z)^{-2} = \left(\Lambda_{u_f(z)} + \Lambda_{u_b(z)}D_N^2\right)^2 = \Lambda_{u_f(z)^2+|u_b(z)|^2} + \Lambda_{2u_f(z)u_b(z)}D_N^2, \tag{4.54}$$

where the juxtaposition of two vectors denotes point-wise multiplication. Finally, the linearity of the trace operator delivers

$$\operatorname{tr}\{M_u(z)\} = \operatorname{tr}\left\{\Lambda_{u_f(z)^2+|u_b(z)|^2}\hat{M}_t(x)\right\} + \operatorname{tr}\left\{\Lambda_{2u_f(z)u_b(z)}\left[D_N^2\hat{M}_t(x)\right]\right\} \tag{4.55}$$

$$= \sum_{m=0}^{N-1}\Big(\left[u_f(z)^2 + |u_b(z)|^2\right]_m [\hat{M}_t(x)]_{m,m} + [2u_f(z)u_b(z)]_m\,[\hat{M}_t(x)]_{N-m,m}\Big) \tag{4.56}$$

with $[\hat{M}_t(x)]_{N,0} \equiv [\hat{M}_c(t)]_{0,0}$. It is important to note that, as desired, Equation (4.56) produces a real output (otherwise we would have a complex rate-loss!) because from (4.42), the fact that M_t is a real matrix, the symmetry of D_N and the identity $D_N^3 = D_N^{\mathrm{H}}$, we have

$$[D_N M_t(x) D_N^{\mathrm{H}}]_{m,l} = [D_N M_t(x)^* D_N^{\mathrm{H}}]_{m,l} = [D_N^{\mathrm{H}} M_t(x) D_N]^*_{m,l} \tag{4.57}$$

$$= \left[D_N^2 \left(D_N M_t(x) D_N^{\mathrm{H}}\right) D_N^2\right]^*_{m,l} = \left[D_N M_t(x) D_N^{\mathrm{H}}\right]^*_{N-m,N-l} \tag{4.58}$$

for all $m,l = 0, 1, \ldots N-1$ with the convention of the indexing mod N. This implies that the sequences $[\hat{M}_t(x)]_{m,m}$ and $[\hat{M}_t(x)]_{N-m,m},\ m = 0, 1, \ldots N-1$ are hermitian symmetric, being thus a real trace in (4.56) obtained[2].

For $L \neq N$ without approximation (case which is only used for small N), we can simply calculate M_t and then $\hat{M}_t$ explicitly, with Equation (4.42), using finally (4.56). Otherwise, for $L = N$ or for the approximated version of $L \neq N$, $\hat{M}_t$ is diagonal (equal to $\Lambda_{N\hat{a}^2}$ for $L = N$ and $\Lambda_{N\hat{\bar{a}}^2}$ for the approximated $L \neq N$), so that the right term in the summand of (4.56) vanishes (remember that $u_b(z,0) = u_b(z,N/2) = 0$) and

$$\mathrm{tr}\{M_u(z)\} = \sum_{m=0}^{N-1} \left[u_f(z)^2 + |u_b(z)|^2\right]_m N\hat{\bar{a}}(z,m)^2 \tag{4.59}$$

(for $N = L$, $\bar{a}$ degenerates in a), so that no explicit N-by-N matrix needs to be computed for the trace of (4.59) in these cases.

For the confirmation that no computation of N-by-N matrices is needed for $L = N$ and for the approximated $L \neq N$, the only thing that is left to do is to express the N-by-N matrices in the second and third terms of (4.50) as cross-diagonal matrices when $\hat{M}_t$ is diagonal. These matrices are M_u, $V^{-1}M_u$ and V^{-2}. Concerning V^{-2}, this task was already done in (4.54). For the other two matrices, we can use the symmetry of $\hat{\bar{a}}^2$, the commutativity of the product of diagonal matrices and property (4.52) to get

$$V(z)^{-1}\Lambda_{N\hat{\bar{a}}(z)^2} = \left(\Lambda_{u_f(z)} + \Lambda_{u_b(z)} D_N^2\right)\Lambda_{N\hat{\bar{a}}(z)^2} = \Lambda_{u_f(z)}\Lambda_{N\hat{\bar{a}}(z)^2} + \Lambda_{u_b(z)} D_N^2 \Lambda_{N\hat{\bar{a}}(z)^2} \tag{4.60}$$

$$= \Lambda_{N\hat{\bar{a}}(z)^2}\Lambda_{u_f(z)} + \Lambda_{D_N^2[N\hat{\bar{a}}(z)^2]}\Lambda_{u_b(z)} D_N^2 = \Lambda_{N\hat{\bar{a}}(z)^2}\left(\Lambda_{u_f(z)} + \Lambda_{u_b(z)} D_N^2\right) \tag{4.61}$$

$$= \Lambda_{N\hat{\bar{a}}(z)^2} V(z)^{-1}. \tag{4.62}$$

We obtain thus, substituting $\hat{M}_t = \Lambda_{N\hat{\bar{a}}^2}$ in (4.44) and using (4.54)

$$M_u(z) = V(z)^{-1}\Lambda_{N\hat{\bar{a}}(z)^2} V(z)^{-1} = \Lambda_{N\hat{\bar{a}}(z)^2} V(z)^{-2} \tag{4.63}$$

$$= \Lambda_{N\hat{\bar{a}}(z)^2[u_f(z)^2+|u_b(z)|^2]} + \Lambda_{N\hat{\bar{a}}(z)^2[2u_f(z)u_b(z)]} D_N^2 \tag{4.64}$$

[2] It is also possible to prove that the diagonal of $\hat{M}_t$ is in fact real (and thus also symmetric), but as that result is not needed to achieve a real trace in (4.56), it will not be proven here.

and

$$\begin{aligned} V(z)^{-1}M_u(z) &= V(z)^{-1}\Lambda_{N\hat{\bar{a}}(z)^2}V(z)^{-2} = \Lambda_{N\hat{\bar{a}}(z)^2}V(z)^{-1}V(z)^{-2} && (4.65)\\ &= \Lambda_{N\hat{\bar{a}}(z)^2}\left(\Lambda_{u_f(z)} + \Lambda_{u_b(z)}D_N^2\right)\left(\Lambda_{u_f(z)^2+|u_b(z)|^2} + \Lambda_{2u_f(z)u_b(z)}D_N^2\right) && (4.66)\\ &= \Lambda_{N\hat{\bar{a}}(z)^2}\left(\Lambda_{u_f(z)[u_f(z)^2+3|u_b(z)|^2]} + \Lambda_{u_b(z)[3u_f(z)^2+|u_b(z)|^2]}D_N^2\right) && (4.67)\\ &= \Lambda_{[N\hat{\bar{a}}(z)^2]u_f(z)[u_f(z)^2+3|u_b(z)|^2]} + \Lambda_{[N\hat{\bar{a}}(z)^2]u_b(z)[3u_f(z)^2+|u_b(z)|^2]}D_N^2. && (4.68) \end{aligned}$$

Joining equations (4.5), (4.6), (2.39), (4.34), (4.37) and (4.50), the rate-loss for the companding scheme of Chapter 3 at high resolution is given (approximately) by

$$\begin{aligned} H(Q_{D,\tilde{F}}) - H(Q_{D,F}) \approx{}& \frac{1}{2N}\sum_{m=0}^{N-1}\mathrm{E}\log_2\left|v_f(Z,m)^2 - |v_b(Z,m)|^2\right| + \frac{1}{N}\mathrm{E}\log_2|\det C(Z)|\\ &- \frac{1}{2N}\mathrm{E}\log_2\det M_t(X) - \frac{\log_2 N}{2}\\ &+ \frac{1}{2}\log_2\Big[\mathrm{E}\ \mathrm{tr}\left\{M_u(Z)\right\} - 2\,\mathrm{E}\ \mathrm{tr}\left\{C(Z)^{-1}\left[H(Z)^{\mathrm{H}}\left(V(Z)^{-1}M_u(Z)\right)A(Z)\right]\right\}\\ &+ \mathrm{E}\ \mathrm{tr}\left\{C(Z)^{-1}\left[H(Z)^{\mathrm{H}}V(Z)^{-2}H(Z)\right]C(Z)^{-\mathrm{T}}\left[A(Z)^{\mathrm{H}}M_u(Z)A(Z)\right]\right\}\Big] \end{aligned} \tag{4.69}$$

$$\begin{aligned} \approx{}& \frac{1}{2N}\sum_{m=0}^{N-1}\mathrm{E}\log_2\left|v_f(Z,m)^2 - |v_b(Z,m)|^2\right| + \frac{1}{N}\mathrm{E}\log_2|\det C(Z)|\\ &- \frac{1}{2N}\sum_{f=0}^{N-1}\mathrm{E}\{\log_2[\hat{\bar{a}}^2(Z,f)]\} - \log_2 N\\ &+ \frac{1}{2}\log_2\Big[\mathrm{E}\ \mathrm{tr}\left\{M_u(Z)\right\} - 2\,\mathrm{E}\ \mathrm{tr}\left\{C(Z)^{-1}\left[H(Z)^{\mathrm{H}}\left(V(Z)^{-1}M_u(Z)\right)A(Z)\right]\right\}\\ &+ \mathrm{E}\ \mathrm{tr}\left\{C(Z)^{-1}\left[H(Z)^{\mathrm{H}}V(Z)^{-2}H(Z)\right]C(Z)^{-\mathrm{T}}\left[A(Z)^{\mathrm{H}}M_u(Z)A(Z)\right]\right\}\Big], \end{aligned} \tag{4.70}$$

where we should use Equation (4.69) for exact values when $L \neq N$, and (4.70) for $L = N$ and for large N in $L \neq N$, taking advantage of the fact that the matrices M_u, $V^{-1}M_u$ and V^{-2} are cross-diagonal to perform matrix multiplications as in the diagram of Figure 3.2(b). Naturally, for $L = N$, M_t and $\bar{a}$ degenerate in M_c and a, respectively.

4.3 Simulations

After calculating the rate-distortion function for the distortion measure in question and the rate-loss incurred by the usage of the suboptimal companding scheme developed in Chapter 3, we are now in condition to simulate the rate-distortion performance of the scheme. In this section, we do this as the main simulation, showing experimentally that the scheme is asymptotically optimal. Furthermore, we also confirm and discuss other theoretical results, namely the approximation of the sensitivity matrix (3.6) by a circulant matrix (Section 3.1), the asymptotic behavior of the rate-distortion function (Subsection 4.1.2) and the validity of the Taylor expansion of Section 3.3. We then proceed explaining a limitation that the distortion measure and the companding scheme have and try to solve that limitation modifying them appropriately. Finally, rate-distortion simulations are repeated and discussed for the modified compander and distortion measure.

4.3.1 Distortion-rate Performance

We will start with the simulations of the distortion-rate performance of the companding scheme of Chapter 3. We implemented the companding scheme based on its Taylor expansion with the parameters of Table 4.1, the distortion-rate function equations without and with approximation of the sensitivity matrix by the circulant matrix of equations (3.54), (3.37), the rate-distortion functions being given by equations (4.7) and (4.8), respectively (using the relation between equations (2.10) and (2.11), we could translate them to the correspondent distortion-rate functions), and the distortion-loss (in dB), without and with approximation, equations (4.69) and (4.70), respectively (and its dependencies), using (2.41) to get the results in terms of the distortion-rate function.

Parameter	Value
Input signal x	Gaussian i.i.d. samples with zero mean, variance σ^2
Power of the input signal σ^2	$0{,}01^2$
Source entropy $h(X)$	$\frac{1}{2}\log_2(2\pi \mathrm{e}\,\sigma^2)$
Vector dimension N	Powers of 2 from 256 to 65536
DFT dimension L	8192 for $N \leq 8192$, N otherwise
Window w	Hamming, $w(n) = 0{,}54 - 0{,}46\cos\left(2\pi\frac{n}{N}\right)$
Sample frequency f_s	48 kHz
Quantizer	$\mathbb{Z}^N$ (component-wise uniform scalar quantization)
Inverse of the quantization scale, $1/s$	5 scales varying exponentially from 10^3 to 10^5
Sphere packing loss SL_D (dB)	$10\log_{10}\left(\frac{2\pi\mathrm{e}}{12}\right) \approx 1{,}5$ dB
Sphere packing loss SL_R (bit/dim)	$\frac{1}{2}\log_2\left(\frac{2\pi\mathrm{e}}{12}\right) \approx 0{,}2546$ bit/dim
Order of Taylor expansion M	3

Table 4.1: Values for the parameters used in the simulations.

We used the calculations for the non-approximated version of the sensitivity matrix for $N \leq 1024$ and the calculations for the approximated version for $1024 \leq N < 8192$. From $N = 8192$ on (inclusive), we used the "approximation" equations since in this case we had $L = N$, delivering thus the equations for the "approximation" exact values (in this case $\bar{M}_c = M_c$ and $\bar{a} = a$). Furthermore, the expected values in the equations were replaced by statistical averages (several realizations of the signal X were emitted), with lower number of realizations for higher vector size N and vice-versa. The reason for decreasing the number of realizations with higher N is that the quantities for which we should estimate the expected value are averages themselves (normalized traces and normalized sums of logarithms of eigenvalues). Assuming that these inner quantities that we are averaging (the eigenvalues and their logarithms) are well behaved, the inner average is consistent, so that these quantities have a low variance for high N. The outer averages (the ones that replace the expectation) reduce the variance furthermore, and thus the lower the N, the heavier that reduction needs to be performed (the higher the needed number of realizations) and vice-versa. In practice, we used on the order of 100 realizations for $N = 256$, of 10 realizations for $N = 1024$ and 1 realization for $N = 8192$ or higher.

As an implementation note, we would like to mention that due to numerical errors on the estimation of the eigenvalues of M_t (necessary for the calculation of the logarithm of its determinant without overloading the numerical representation), its weakest values were considered numerical noise (see Figure 4.1 for an example). To avoid this problem, the eigenvalues of M_t and the ones of the approximation $\bar{M}_c$ (which are $N\hat{\bar{a}}(x, f)$, $f = 0, 1, \ldots, N-1$) were sorted and the noisy eigenvalues of M_t were substituted by the correspondent ones in $\bar{M}_c$. The decision on which eigenvalues were noisy was made on basis of a threshold on the index of 41 kHz $\cdot N/f_s$; the index values larger than the threshold were replaced. Note that sorting does not affect the result of the previous calculations, because the relevant equations are written in terms of sums of functions of eigenvalues, and the sum is a commutative operation. Also an implementation detail, when implementing Equation (4.56), as the values of u_f increase abruptly in the high and in the very low frequency range (u_f is given in Equation (3.187) and v_f decays to 0 rapidly in those ranges), the result of the multiplication on the left side of (4.56) yields a result which is on a much higher order of magnitude in that range than outside the range. Again, this problem was solved with the substitution by the approximated values, with the thresholds 150 Hz $\cdot N/f_s$ and 13 kHz $\cdot N/f_s$ in the low and high frequencies, respectively.

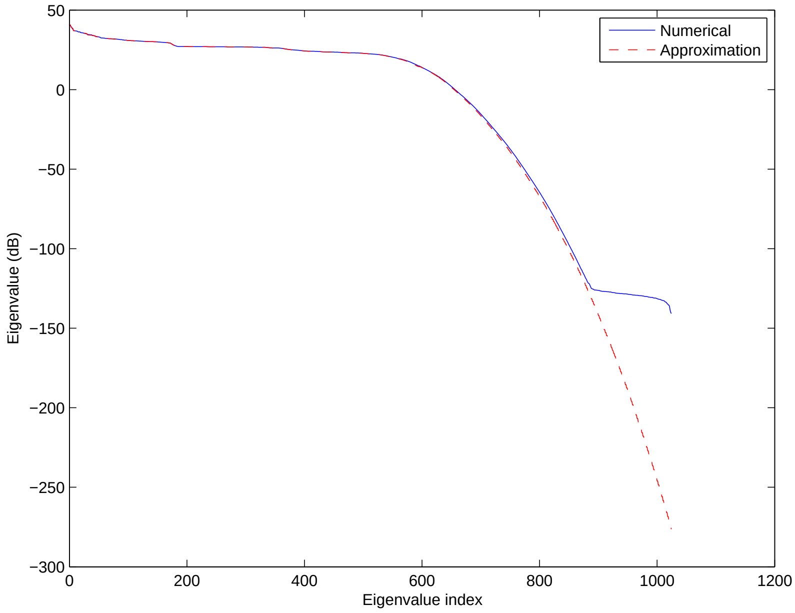
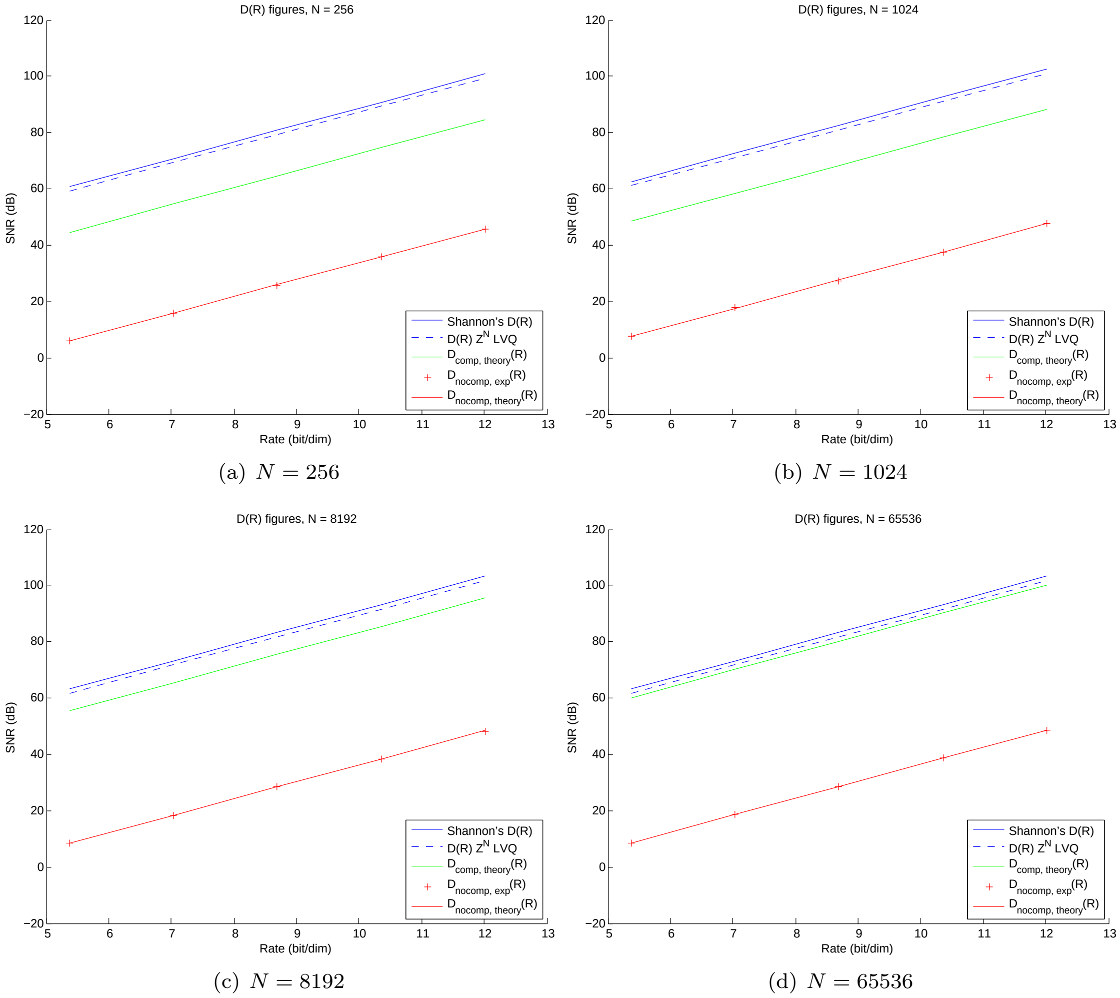


Figure 4.1: Sorted eigenvalues of the Toeplitz matrix M_t calculated numerically and using the approximation by the circulant matrix $\bar{M}_c$.

The results of the main simulation, several distortion-rate figures for varying vector dimension, are shown in Figure 4.2 (only the results corresponding to the vector dimensions $N \in \{256, 1024, 8192, 65536\}$ are shown). In Figure 4.3 we show the distortion-rate figures for $N = 1024$ with approximation overlaid with the ones without approximation of the Toeplitz M_t by the circulant $\bar{M}_c$.



(a) $N = 256$

(b) $N = 1024$

(c) $N = 8192$

(d) $N = 65536$

Figure 4.2: Distortion-rate performance of the companding scheme for varying N. Distortions are in terms of the SNR in dB $(10 \log_{10}(\sigma^2/D))$. The blue line represents the Shannon's distortion-rate function. The blue dotted line gives the best achievable $\mathbb{Z}^N$ lattice vector quantizer (LVQ) distortion-rate performance. The green and red lines give the performance of the companding scheme and of the identity compander $F(x) = F^{-1}(x) = x$. The red crosses are results obtained from quantizing the source directly (i.e., using the identity compander), and calculating the rate and distortion values directly.

The vertical axis represents the signal-to-noise ratio (SNR) in dB, i.e., $D_{\text{dB}} = 10 \log_{10}(\sigma^2/D)$, and the horizontal axis represents the rate per dimension R. The blue line depicts the Shannon's distortion-rate function of (4.7) or (4.8), $D(R)$. The blue dotted line is the best ever achievable with a $\mathbb{Z}^N$ lattice vector quantizer (LVQ). It is given by the $D(R)$ function minus the sphere packing loss in dB. The green line, given by the $D(R)$ function minus the sphere packing loss minus the rate-loss of Equation (4.69) or (4.70), is the performance that we have theoretically upon the usage of the developed companding scheme (remember that the optimal compander would run on the blue dotted line, if it would exist). The red line, given by the $D(R)$ function minus the sphere packing loss minus the rate-loss for the identity compander of Equation (4.24) or (4.25), is the theoretical performance without compander. Finally, the red crosses represent the measured $D(R)$ performance without compander, obtained from quantizing the source directly, measuring the distortion with Equation (2.42) and estimating the rate at the output of the entropy coder with the high-resolution approximation [23]

$$H(Q_{D(s),\mathrm{I}}) \approx h(X) - \frac{1}{N}\log_2(\mathrm{Vol}(s\Lambda)) \tag{4.71}$$

$$= h(X) - \frac{1}{N}\log_2(s^N \mathrm{Vol}(\mathbb{Z}^N)) = h(X) - \log_2 s, \tag{4.72}$$

where s is the quantization scale factor and Vol denotes the hyper-volume of the basic cell of the quantizer, which is 1 for the $\mathbb{Z}^N$ lattice. For Figure 4.3, the blue, dotted blue, dotted green and dotted red lines and the red crosses have the same meaning as the lines in Figure 4.2 (the red and green line not being dotted), and the light-blue, dotted light-blue, dark green and brown lines correspond to the blue, dotted blue, green and red lines, respectively, but for the calculations with approximation of the Toeplitz matrix M_t by the circulant matrix $\bar{M}_c$.

Figure 4.2 corroborates the theoretical considerations on asymptotical optimality. Indeed, the theoretical performance of the companding scheme gets very close to the best possible, the one of the optimal scheme with the $\mathbb{Z}^N$ LVQ. For a higher detail, see Table 4.2, which shows the distortion loss of the companding scheme for varying dimension N (the distances between the performance of the compander and the performance of an optimal scheme using a $\mathbb{Z}^N$ LVQ, i.e., between the green and blue dotted line of Figure 4.2). To get the rate-loss, it suffices to use the proportionality rule that a gain of 6 dB corresponds to an additional rate of 1 bit. Note that $N = 1024$ is the vector size which corresponds to the frame-size used in practice (N/f_s = 21,3 ms); (much) lower or higher values of N do not correspond to the frequency and time resolution of the human ear, respectively. Furthermore, for $N/f_s \gg 20$ ms, the audio signal cannot be considered stationary.

N	256	512	1024	2048	4096	8192	16384	32768	65536
Distortion Loss (dB)	14,6	14,0	12,6	11,0	8,65	6,27	4,28	2,77	1,57

Table 4.2: Distortion Loss of the companding scheme for varying N.

We can also see through Figure 4.2 that the identity compander does not achieve asymptotic optimality, as explained in Subsection 4.2.1. This motivates the usage of the compressor: in [28], a companding scheme was developed for a different distortion measure, and the authors concluded that the performance of the identity compander is so close to the optimal performance that it is not worth to do companding or normalization by perceptual weights; the identity compander could be

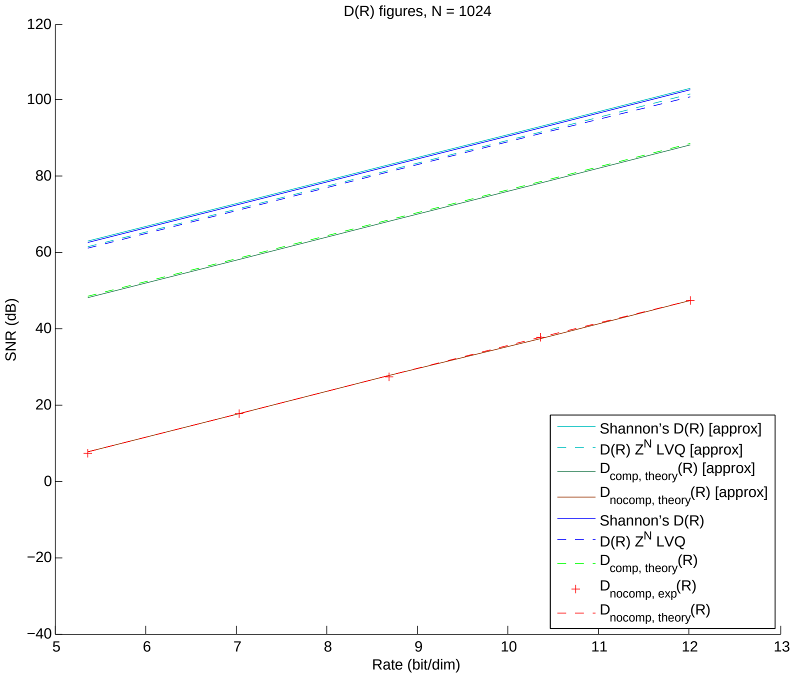


Figure 4.3: Distortion-rate performance of the companding scheme for $N = 1024$. Comparison of the circulant matrix approximation with the exact values. See Figure 4.2 for the meaning of the common colors (the dotted red and green here are exactly the same as the solid red and green in Figure 4.2, respectively). The common colors between the two figures refer to the non-approximated version. The light-blue, dotted light-blue, dark green and brown lines correspond to the blue, dotted blue, green and red lines, respectively, but for the approximated version.

directly used without a big penalty in performance. For our distortion measure, this is obviously not the case, since the rate-loss incurred by not using any companding scheme is approximately 9 bit/dim (distortion loss of 54 dB).

Additionally, note that Figure 4.3 confirms the approximations of the matrix M_t for N sufficiently large. Indeed, the approximated and non-approximated versions are on top of each other.

Although not visible at first sight, the distortion-rate functions of Figure 4.2 (and corresponding best LVQ optimal performances) reach a steady offset when $N \to \infty$. That can be seen more clearly through Table 4.3, which shows the additional term of the distortion-rate function with respect to the Shannon lower bound ($1/(2N)$ $\mathrm{E}[\log_2 \det M(X)]$). This excess term converges with increasing N, which is in line with the theoretical results of Subsection 4.1.2; the remaining terms of the distortion-rate function (4.7) or (4.8) with (2.11) only depend on normalized values, thus not changing with N.

N	256	512	1024	2048	4096	8192	16384	32768	65536
Excess	-4,74	-4,93	-5,01	-5,13	-5,13	-5,14	-5,13	-5,14	-5,14

Table 4.3: Excess term of $D(R)$ with respect to the Shannon lower bound, $1/(2N)$ $\mathrm{E}[\log_2 \det M(X)]$.

The attentive reader probably noticed an important lack in the graphs of Figure 4.2: the green crosses, i.e., the measured performance values with companding scheme were missing. For producing such results, we would need to measure the distortion between the signals before compression and after expansion (which is not difficult; we just have to implement the whole chain and use (2.42)), and measure the entropy at the output of the quantizer, using the approximation (4.71) with $F(X)$ instead of X. The problem is in this last procedure: as the differential entropy of (4.71) refers to a vector size N which is on the order of hundreds or thousands, and as the components of the vector $F(X)$ are not independent, to estimate $h[F(X)]$, we would need to gather an extremely high number of (N-dimensional) samples of $F(X)$ for the estimation of its probability density function (through the calculation of its N-dimensional histogram), and then of its entropy. That process is thus not computationally feasible. There are other methods for estimating the entropy, which are not based on histogram computations (the so-called *plug-in estimators*), such as sample-spacings estimators [6, 35] and nearest-neighbors estimators [6]. Another way to estimate $h[F(X)]$ would be to use the random variable substitution rule to find out the probability density function of $F(X)$ in terms of the one of X, and then to calculate the differential entropy using Equation (2.9), (with $F(X)$ instead of X). Nevertheless, the estimation of $h[F(X)]$ is out of the scope of this thesis, being left for future work.

4.3.2 Validity of the Taylor Expansion

To test how good the approximation of the original compressor of Section 3.2 by the Taylor expansion of Section 3.3 is, and what order of M we must use to get a good approximation, we computed the contribution of the 0^{th} to the 6^{th} order terms of the signal dependent gain Γ of Equation (3.89) for the vector size most useful in practice, $N = 1024$. Remember that the compressor, Equation (3.90), essentially point-wise multiplies the windowed input in the frequency domain by the gain. The contributions can be seen in Figure 4.4, depicted in double-logarithmic scale. The figure shows that already for $N = 1024$, if we use $M = 0$, we get only a small difference between the original and the Taylor expanded compressor. If we want a finer detail we can use higher M, but the order of units is more than enough. Indeed, the distance between the 0^{th} and 1^{st} order terms is at least 15,7 dB, which means that the 1^{st} order term has values that are attenuated at least 6,1 times with respect to the values of the 0^{th} order term. For the 2^{nd} order term this distance is 23,9 dB (15,7 times), and for the 6^{th} order 42,8 dB (138 times).

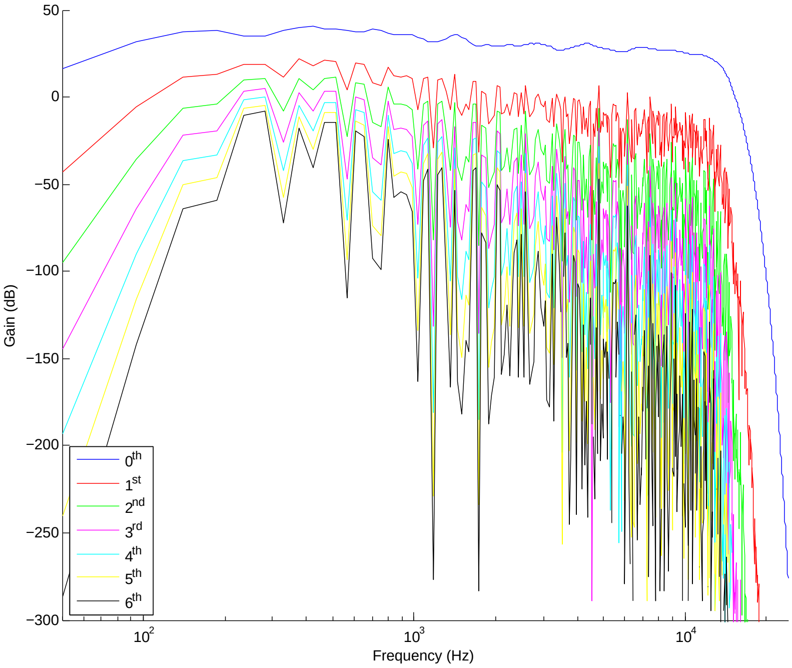
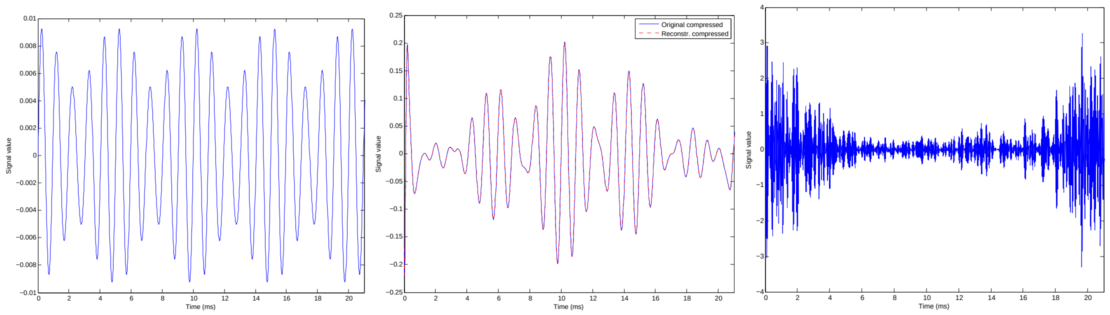


Figure 4.4: Contribution of the j^{th} order term of the Taylor expansion of the signal-dependent Γ, $j = 0, 1, \ldots, 6$.

4.3.3 Limitations

Figure 4.4 shows an important limitation that the developed multidimensional companding scheme exhibits: as the gain Γ decreases abruptly with increasing frequency, the inversion of the compressor, although mathematically possible (at least for $M = 0$; see Subsection 3.5.1), is a numerically badly conditioned problem. Using the rough rationale based on Equation (3.170) that the expander multiplies the incoming signal point-wise by the inverse of Γ, we can see through Figure 4.4 that the quantization noise present in the input of the expander will be heavily amplified in the high frequencies (imagine the characteristics upside-down, since the inverse in the linear domain corresponds to the symmetric in the log domain). The same happens for the very low frequencies. In practice, this means that two very different signals in the non-compressed domain will have very near or even equal compressed signals, up to numerical noise. This situation is depicted in Figure 4.5, where we applied the multidimensional companding chain (Figure 1.2) with a small quantization step size of $s = 10^{-5}$ to an input signal given by two sinusoids (frequencies 1 kHz and 1,2 kHz at levels 50 dBSPL and 40 dBSPL, respectively), and observed the original, the expanded and the compressed and re-compressed expanded signal.

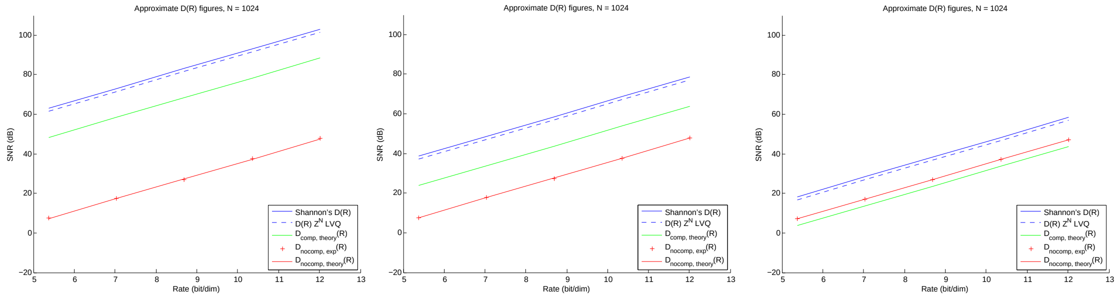


(a) Original signal (b) Compressed domain (c) Expanded signal

Figure 4.5: Numerical bad conditioning of the inversion of the compressor.

To deal with this limitation, we can crop the gain Γ and the masking threshold $\hat{\hat{a}}$ of the distortion measure (2.42), i.e., if the frequency is below a certain threshold in the low-frequency range, we use the boundary value (the value on the threshold) of the gain/masking threshold, and if the frequency is above another threshold, in the high-frequency range, we use the boundary value there. Between the two thresholds, the original gain/masking threshold are conserved. For $N = 1024$, the distortion-rate performance of the compressor was simulated again cropping Γ and $\hat{\hat{a}}$ only in the high frequency range at $\{24, 18, 14\}$ kHz (the first case corresponds to not cropping at all, since $f_s/2 = 24$ kHz), and then cropping in the low frequency range at $\{50, 100\}$ Hz for no high-frequency cropping and for high-frequency cropping at 18 kHz. The results are shown in figures 4.6 and 4.7, respectively.

We observe a degradation of performance for decreasing threshold of the high-frequency crop: the distortion-rate function, the $\mathbb{Z}^N$ best distortion-rate performance and the distortion-rate performance of the compander all sink in the graphs, i.e., for the same rate, the best achievable and the actual distortions increase (the SNR decreases). At a certain point (in the figure, at 14 kHz) the performance of the compander sinks below the performance of not doing anything, i.e., of using the identity compander $F(x) = x$. For the low-frequency cropping case no performance loss is obtained, as the compander curves lie on top of each other. For practical numerical well condi-

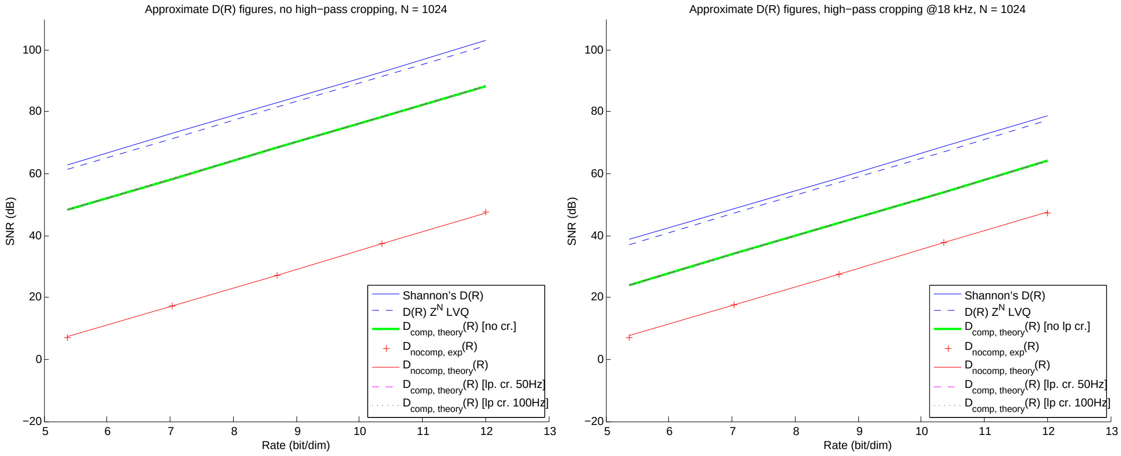


(a) 24 kHz (no cropping)

(b) 18 kHz

(c) 14 kHz

Figure 4.6: Distortion-rate performances for cropping in the high-frequency range at 24, 18 and 14 kHz.



(a) no cropping

(b) high-frequency cropping at 18 kHz

Figure 4.7: Distortion-rate performances for cropping in the low-frequency range at 50 and 100 Hz, for no high-frequency cropping and for cropping at 18 kHz.

tioning, a cropping in the low frequencies at 50 Hz and in the high frequencies at 18 kHz suffices. The graph of Figure 4.7(b) (with the dotted magenta line, corresponding to the low frequency cropping at 50 Hz) is thus the one that depicts the performance of the scheme in a practically situation.

Chapter 5

Conclusion

5.1 Summary

In this thesis paper, a multidimensional companding scheme was developed for the perceptual distortion measure [59]. Although the developed scheme is suboptimal, it was proven that no optimal scheme exists under a wide range of solutions for the optimality condition (2.19), and that, nevertheless, the conceived scheme reached optimality asymptotically, i.e., with increasing vector dimension N.

The scheme was developed in the frequency domain; in its most simple form, the compressor windows the input signal, applies a Discrete Fourier Transform (DFT), multiplies the input signal by the square-root of the inverse of its masking threshold (the square-root of Equation (2.44)), and then goes back into the time domain with the inverse DFT. This process is exactly the same as the process of normalization by perceptual weights, represented in Equation (2.62), but with the advantage that no perceptual weights have to be transmitted through the channel.

At the receiver, an algorithm based on numerical methods, the expander, has to be run. The expander does not depend on the perceptual weights and it can be run using the previous audio frame, already available at the decoder, as its initial condition.

The theoretical results and assumptions were corroborated with simulations.

5.2 Future Work

As in every work, due to restrictions of all types, it was not possible to complete a thorough research on this theme, and some topics were consequently left open. We would like to point out some of those topics, as an indication for possible future work.

On the theoretical side, it would be interesting to explore fully the conditions in which an optimal companding scheme exists, both in general and for the particular sensitivity matrix derived in this work (see Section 3.1 and the optimality condition, Equation (2.19)). It was proven that, given a certain square-root of the sensitivity matrix, the Schwarz' theorem could be applied to it to obtain a necessary and sufficient condition for the existence of an optimal scheme (Subsection 2.2.4). These results were extended for other square-roots, given in terms of the original square-root through a left multiplication by an orthogonal matrix. The extension only covered the case in which the orthogonal matrix did not depend on the input signal. It would be thus interesting to explore the (more complex) case of orthogonal matrices dependent on the input signal. This case covers all possible square-roots of the sensitivity matrix, as explained in Subsection 2.2.4.

Another area where theoretical work could be done is on proving (or disproving) the invertibility of the constructed compressor almost everywhere (with probability 1) when a Taylor expansion of order $M > 0$ (or even $M = \infty$) is used. In this thesis, such a proof was only delivered for $M = 0$ (Subsection 3.5.1). Although this is an important theoretical step to use the companding scheme based on this compressor for $M > 0$ (independently of the form of the expander), we have seen in the simulations (Section 4.3) that $M = 0$ forms the most interesting case, since the other terms of the Taylor expansion are negligible with respect to the one for $M = 0$. It is thus a wise suggestion that future work on this theme should rely on the compressor with a Taylor expansion of order $M = 0$.

A radical change on the work done here, but still in the framework of multidimensional companding for this distortion measure, would be trying other compressors. One possibility in that direction is trying to apply the work of Heusdens et al. [28] for this distortion measure with the substitution explained in Subsection 2.3.3. It seems to the author that the way to go here would be applying a multi-rate technique to downsample the signal of the work [28] of size NP down to the size N, and then upsample it back at the receiver. We would have to investigate in which conditions no loss of information occurs, and how exactly the distortion measure (2.42) maps into the one of Heusdens (Equation (2.61)) in that case. A less radical change in the framework of design alterations would be to develop a compressor using a different square-root of the sensitivity matrix, integrating the appropriate equations in it (and also developing the correspondent expander, obviously).

Coming back again to the work of this thesis paper, one point that was missing in the simulations was the comparison of the theoretical distortion-rate performance of the companding scheme with practical "measured" values (Section 4.3). For doing that comparison, we would have to find an efficient way to estimate the differential entropy of an N-dimensional source (with N large) using a tolerable number of vector samples. An additional point that was not done at the simulations and that would be useful to get a global fair overview of the scheme, was to compare the (distortion-rate) performance of the companding scheme with the performance of a scheme based on the perceptual weighting of Equation (2.62).

Regarding the bad conditioning of the inversion of the compressor, other solutions of the problem could be explored, such as the usage of another filterbank h_i which still correlates reasonably with the behavior of the inner ear but that does not exhibit the problem of a (too) quickly decaying inverse of the masking threshold for very low and high frequencies.

As a final indication for future work, note that the development of the expander is not yet complete. As explained in Section 3.5, the expander works by performing one fixed point iteration and fine-tuning the result using Broyden's method. In the first place, there are computational complexity issues here: although the fixed point iteration is a light computation (it has the same complexity as the calculation of the compressor), the fine-tuning process is still expensive due to the compulsory calculation of the Jacobian matrix (or elements of it) in the first iteration. One possibility for future work is thus the research on less computationally expensive numerical methods. In the second place, independently of the numerical method used, a very important condition to check is that the numerical method converges. Indeed, it is known that the Newton and Broyden methods converge [32] when the initial guess is "sufficiently close" to the root of the equation to solve but, in quantitative terms, how close should the initial estimate be, and how can we find an initial estimate so that the method converges always? Of course another possibility for future work would be to try to find an analytical inverse of the compressor, at least for $M = 0$. This would in principle solve the complexity issues and the problem of doubtful convergence.

Bibliography

[1] ISO/IEC 11172-3:1993. *Coding of Moving Pictures and Associated Audio for Storage at up to about 1.5 Mbit/s, part 3: Audio*, 1993. ISO/IEC International Standard.

[2] ISO/IEC 14496-3:2001. *Coding of audio-visual objects – Part 3: Audio (MPEG-4 Audio Edition 2001)*, 2001. ISO/IEC International Standard.

[3] ISO/IEC 14496-3:2001. *HILN, Harmonic and Individual Lines Plus Noise*, 2001. ISO/IEC International Standard.

[4] ISO/IEC 14496-3:2001/Amd 2:2004. *Parametric coding for High-Quality Audio*, 2004. ISO/IEC International Standard.

[5] T. M. Apostol. *Calculus, Vol. 2: Multi-Variable Calculus and Linear Algebra with Applications*, volume 2. Wiley, second edition, 1969.

[6] J. Beirlant, E. J. Dudewicz, L. Gyorfi, and E. C. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6:17 – 39, 1997.

[7] W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27:446 – 472, July 1948.

[8] C. A. Berenstein and R. Gay. *Complex Variables, An Introduction.* Springer-Verlag, 1991.

[9] T. Berger. *Rate distortion theory, a mathematical basis for data compression.* Prentice-Hall, 1971.

[10] J. Bucklew. Companding and random quantization in several dimensions. *IEEE Trans. on Information Theory*, 27:207 – 211, March 1981.

[11] J. Bucklew. A note on optimal multidimensional companders. *IEEE Trans. on Information Theory*, 29:279, March 1983.

[12] J. Y. Choi and J. D. H. Smith. On the unimodilty and combinatorics of bessel numbers. *Discrete Math.*, 264:45 – 53, 2003.

[13] J. B. Conway. *Functions of One Complex Variable I.* Springer-Verlag, second edition, 1978.

[14] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, July 2006.

[15] T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the effective signal processing in the auditory system. i. model structure. *J. Acoust. Soc. Amer.*, 99(6):3615 – 3622, June 1996.

[16] R. M. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge, UK, 2002.

[17] B. Edler and G. Schuller. Audio coding using a psychoacoustic pre- and post-filter. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 881 – 884, June 2000.

[18] Bernd Edler and Heiko Purnhagen. Parametric audio coding. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Mohonk, New Paltz*, pages 31–34, 2000.

[19] R. Heusdens et al. Bit-rate scalable intraframe sinusoidal audio coding based on rate-distortion optimization. *J. Audio Eng. Soc.*, 54(3):167 – 188, March 2006.

[20] A. Gersho. Asymptotically optimal block quantization. *IEEE Trans. Information Theory*, 25(4):373 – 380, July 1979.

[21] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1992.

[22] V. K. Goyal. Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine*, 18(5):74 – 94, Sept. 2001.

[23] R. Gray and D. Neuhoff. Quantization. *IEEE Trans. on Information Theory*, 44(6):2325 – 2383, 1998.

[24] R. M. Gray. *Source Coding Theory.* Kluwer Academic Publishers, 1990.

[25] R. M. Gray. *Toeplitz and Circulant Matrices: A Review.* Now Publishers, Norwell, Massachusetts, 2006.

[26] G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, third edition, 2001.

[27] D. A. Harville. *Matrix Algebra From a Statistician's Perspective.* Springer-Verlag, 1997.

[28] R. Heusdens, W. B. Kleijn, and A. Ozerov. Entropy-constrained high-resolution lattice vector quantization using a perceptually relevant distortion measure. *Proc. of the 2007 Asilomar Conference on Signals, Systems and Computers*, pages 2075–2079, Nov. 2007.

[29] R. Horn and C. Johnson. *Matrix Analysis.* Cambridge University Press, 1990.

[30] D. G. Jeong and J. D. Gibson. Image coding with uniform and piecewise-uniform vector quantizers. *IEEE Trans. on Image Processing*, 4(2), Feb. 1995.

[31] J. D. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2):314 – 323, Feb. 1988.

[32] C. T. Kelley. *Solving Non-linear Equations with Newton's Method.* SIAM, 2003.

[33] D. Kincaid and W. Cheney. *Numerical Analysis: Mathematics of Scientific Computing.* Brooks/Cole, third edition, 2002.

[34] P. Korten, J. Jensen, and R. Heusdens. High resolution spherical quantization of sinusoidal parameters using a perceptual distortion measure. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 177 – 180, March 2005.

[35] E. G. Learned-Miller. A new class of entropy estimators for multi-dimensional densities. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, volume 3, pages 297 – 300, April 2003.

[36] T. Linder and R. Zamir. On the asymptotic tightness of the shannon lower bound. *IEEE Trans. Inf. Theory*, 40(6):2026 – 2031, Nov. 1994.

[37] T. Linder and R. Zamir. High-resolution source coding for non-difference distortion measures: The rate-distortion function. *IEEE Transactions on Information Theory*, 45(2):533–547, March 1999.

[38] T. Linder, R. Zamir, and K. Zeger. High-resolution source coding for non-difference distortion measures: Multidimensional companding. *IEEE Transactions on Information Theory*, 45(2):548–561, March 1999.

[39] T. Linder and K. Zeger. Asymptotic entropy-constrained performance of tesselating and universal randomized lattice quantization. *IEEE Trans. on Information Theory*, 40(2):575 – 579, March 1994.

[40] H. S. Malvar. *Signal Processing with Lapped Transforms.* Artech House, 1992.

[41] V. Melkote and K. Rose. A modified distortion metric for audio coding. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) 2009*, April 2009.

[42] P. Moo and D. Neuhoff. Optimal compressor functions for multidimensional companding. In *Proc. IEEE Int. Symp. Information Theory, Ulm, Germany*, page 515, 1997.

[43] O. Niemeyer and B. Edler. Efficient coding of excitation patterns combined with a transform audio coder. In *Proc. of the 118th AES Conv., Barcelona, Spain*, May 2005.

[44] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer-Verlag, 1999.

[45] A. Oppenheim, R. Schafer, and J. Buck. *Discrete-Time Signal Processing.* Prentice Hall, 1999.

[46] T. Painter and A. Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4):451 – 515, April 2000.

[47] P. F. Panter and W. Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. In *Proc. of the IRE*, volume 39, no. 1, pages 44 – 48, 1951.

[48] J. H. Plasberg and W. B. Kleijn. The sensitivity matrix: Using advanced auditory models in speech and audio processing. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(1):310 – 319, Jan. 2007.

[49] S. S. Pradhan, R. Puri, and K. Ramchandran. n-channel symmetric multiple descriptions – part i: (n, k) source-channel erasure codes. *IEEE Trans. on Information Theory*, 50(1):47 – 61, Jan. 2004.

[50] H. Purnhagen. Advances in parametric audio coding. In *Proc. 1999 IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics, New Paltz, New York, USA*, pages W99–1 – W99–4, 1999.

[51] W. R. Gardner B. D. Rao. Theoretical analysis of the high-rate vector quantization of lpc parameters. *IEEE Trans. on Speech and Audio Processing*, 3(5):367 – 381, Sept. 1995.

[52] D. J. Sakrison. The rate distortion function of a gaussian process with a weighted square error criterion. *IEEE Trans. Inf. Theory*, IT-14:506 – 508, May 1968.

[53] Jonas Samuelsson. Multidimensional companding quantization of the gaussian source. *IEEE Trans. on Information Theory*, 49(5), May 2003.

[54] V. Scheidemann. *Introduction to Complex Analysis in Several Variables.* Birkhäuser-Verlag, 2005.

[55] S. Simon. On suboptimal multidimensional companding. In *Proc. Data Compression Conf.*, pages 438 – 447, 1998.

[56] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault. Subjective evaluation of state-of-the-art 2-channel audio codecs. *J. Audio Eng. Soc.*, 46(3):164 – 176, March 1998.

[57] A. D. Subramaniam, W. R. Gardner, and B. D. Rao. Low-complexity source coding using gaussian mixture models, lattice vector quantization, and recursive coding with application

to speech spectrum quantization. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2), March 2006.

[58] R. Vafin and W. Kleijn. Entropy-constrained polar quantization and its application to audio coding. *IEEE Trans. on Speech and Audio Processing*, 13(2):220 – 232, March 2005.

[59] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. Jensen. A perceptual model for sinusoidal audio coding based on spectral integration. *EURASIP Journal on Applied Signal Processing*, 2005(9):1292–1304, June 2005.

[60] M. Vetterli, J. Kavacevic, and V. Goyal. The world of fourier and wavelets. Published in `http://www.fourierandwavelets.org/`, 2008.

[61] E. Vincent and M. D. Plumbley. Low bit-rate object coding of musical audio using bayesian harmonic models. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(4):1273 – 1282, May 2007.

[62] R. Gray Y. Yamada, S. Tazaki. Asymptotic performance of block quantizers with difference distortion measures. *IEEE Trans. on Information Theory*, 26(1):6 – 14, Jan. 1980.

[63] P. Zador. Asymptotic quantization of continuous random variables. Technical report, Bell Laboratories, 1966. unpublished memorandum.

[64] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Heidelberg: Springer, second edition, 1999.

[65] J. Østergaard. *Multiple-Description Lattice Vector Quantization*. PhD thesis, Technical University of Delft, 2007.

Appendix A

Proofs for Identities Used in the Document

A.1 The Binomial Theorem for the Derivative of a Product

Theorem 1. *Let*

$$a, b : \Omega \subseteq \mathbb{R} \to \mathbb{R} \tag{A.1}$$

be two real-valued real functions defined in an open set, Ω, which are k-times differentiable in it, $k = 1, 2, \ldots$ Then,

$$\begin{aligned} c &: \Omega \subseteq \mathbb{R} \to \mathbb{R} \\ c(x) &= a(x)\,b(x) \quad \forall x \in \Omega \end{aligned} \tag{A.2}$$

is k-times differentiable with

$$c^{(k)}(x) \overset{\text{def}}{=} \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left[a(x)b(x)\right] = \sum_{n=0}^{k} \binom{k}{n} \frac{\mathrm{d}^n a}{\mathrm{d}x^n}(x) \frac{\mathrm{d}^{k-n} b}{\mathrm{d}x^{k-n}}(x), \quad \forall x \in \Omega. \tag{A.3}$$

Proof. The proof is made by induction. First, note that, for $k = 1$,

$$c^{(1)}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[a(x)b(x)\right] = \sum_{n=0}^{1} \binom{1}{n} \frac{\mathrm{d}^n a}{\mathrm{d}x^n}(x) \frac{\mathrm{d}^{1-n} b}{\mathrm{d}x^{1-n}}(x) = a(x)\frac{\mathrm{d}b}{\mathrm{d}x}(x) + \frac{\mathrm{d}a}{\mathrm{d}x}(x)b(x) \tag{A.4}$$

is a simple application of the rule for the derivative of the product of two differentiable functions.

Now assume that statement (A.3) is valid for a certain order k_0, $k_0 = 0, 1, \dots, k-1$. As

$$c^{(k_0)}(x) = \frac{\mathrm{d}^{k_0}}{\mathrm{d}x^{k_0}}\left[a(x)b(x)\right] = \sum_{n=0}^{k_0} \binom{k_0}{n} \frac{\mathrm{d}^n a}{\mathrm{d}x^n}(x) \frac{\mathrm{d}^{k_0-n} b}{\mathrm{d}x^{k_0-n}}(x) \tag{A.5}$$

is a sum of a scaled product of functions which are differentiable (in particular, a and b are differentiable $k \geq k_0 + 1$ times), $c^{(k_0)}$ is itself differentiable and thus c is differentiable $k_0 + 1$ times with

$$c^{(k_0+1)}(x) = \frac{\mathrm{d}\, c^{(k_0)}}{\mathrm{d}x}(x) = \sum_{n=0}^{k_0} \binom{k_0}{n} \left(\frac{\mathrm{d}^{n+1} a}{\mathrm{d}x^{n+1}}(x) \frac{\mathrm{d}^{k_0-n} b}{\mathrm{d}x^{k_0-n}}(x) + \frac{\mathrm{d}^n a}{\mathrm{d}x^n}(x) \frac{\mathrm{d}^{k_0-n+1} b}{\mathrm{d}x^{k_0-n+1}}(x) \right). \tag{A.6}$$

This followed directly from the usage of the product rule and of the sum rule in differentiation.

Substitution of variables on the right term and a bit of algebraic work yields

$$\begin{aligned} c^{(k_0+1)}(x) &= \sum_{n=0}^{k_0} \binom{k_0}{n} \frac{\mathrm{d}^{n+1} a}{\mathrm{d}x^{n+1}}(x) \frac{\mathrm{d}^{k_0-n} b}{\mathrm{d}x^{k_0-n}}(x) + \sum_{n=-1}^{k_0-1} \binom{k_0}{n+1} \frac{\mathrm{d}^{n+1} a}{\mathrm{d}x^{n+1}}(x) \frac{\mathrm{d}^{k_0-n} b}{\mathrm{d}x^{k_0-n}}(x) && \text{(A.7)} \\ &= \sum_{n=0}^{k_0-1} \left[\binom{k_0}{n} + \binom{k_0}{n+1} \right] \frac{\mathrm{d}^{n+1} a}{\mathrm{d}x^{n+1}}(x) \frac{\mathrm{d}^{k_0-n} b}{\mathrm{d}x^{k_0-n}}(x) + a(x) \frac{\mathrm{d}^{k_0+1} b}{\mathrm{d}x^{k_0+1}}(x) + \frac{\mathrm{d}^{k_0+1} a}{\mathrm{d}x^{k_0+1}}(x) b(x). && \text{(A.8)} \end{aligned}$$

Using the recursive definition of the binomial coefficients,

$$\binom{k_0+1}{n+1} = \binom{k_0}{n} + \binom{k_0}{n+1}, \quad n = 0, 1, \dots, k_0 - 1,\ k_0 = 1, 2, 3, \dots, \tag{A.9}$$

we obtain

$$\begin{aligned} c^{(k_0+1)}(x) &= \sum_{n=0}^{k_0-1} \binom{k_0+1}{n+1} \frac{\mathrm{d}^{n+1} a}{\mathrm{d}x^{n+1}}(x) \frac{\mathrm{d}^{k_0-n} b}{\mathrm{d}x^{k_0-n}}(x) + a(x) \frac{\mathrm{d}^{k_0+1} b}{\mathrm{d}x^{k_0+1}}(x) + \frac{\mathrm{d}^{k_0+1} a}{\mathrm{d}x^{k_0+1}}(x) b(x) && \text{(A.10)} \\ &= \sum_{n=1}^{k_0} \binom{k_0+1}{n} \frac{\mathrm{d}^n a}{\mathrm{d}x^n}(x) \frac{\mathrm{d}^{k_0+1-n} b}{\mathrm{d}x^{k_0+1-n}}(x) + a(x) \frac{\mathrm{d}^{k_0+1} b}{\mathrm{d}x^{k_0+1}}(x) + \frac{\mathrm{d}^{k_0+1} a}{\mathrm{d}x^{k_0+1}}(x) b(x) && \text{(A.11)} \\ &= \sum_{n=0}^{k_0+1} \binom{k_0+1}{n} \frac{\mathrm{d}^n a}{\mathrm{d}x^n}(x) \frac{\mathrm{d}^{k_0+1-n} b}{\mathrm{d}x^{k_0+1-n}}(x), && \text{(A.12)} \end{aligned}$$

which proves that (A.3) is also valid for order $k_0 + 1$. Applying the previous considerations recursively for $k_0 = 1, 2, 3, \dots, k-1$ yields the complete proof of the theorem.

□

A.2 Determinant and Inverse of a Cross-diagonal Matrix

Theorem 2. *Define as Λ_v the diagonal matrix whose diagonal elements are the elements of the vector v. Furthermore, let $f = [f(0), f(1), f(2), \ldots, f(N-1)]^{\mathrm{T}}$ and $b = [b(0), b(1), b(2), \ldots, b(N-1)]^{\mathrm{T}}$ be two N-dimensional complex vectors, $N = 2, 4, 6, \ldots$*[1]*, where $b(0) = b(N/2) = 0$, and let*

$$X = \Lambda_f + \Lambda_b D_N^2 \tag{A.13}$$

be a cross-diagonal matrix, where D_N^2 is the matrix operator of Equation (3.60)*. Then we have*

$$\det X = f(0) f\left(\frac{N}{2}\right) \prod_{m=1}^{\frac{N}{2}-1} \left[f(m) f(N-m) - b(m) b(N-m)\right] \tag{A.14}$$

with $\prod_{m=1}^{0} \equiv 1$. Furthermore, if none of the factors of (A.14) *vanishes, then X is invertible with*

$$X = \Lambda_\phi + \Lambda_\beta D_N^2, \tag{A.15}$$

where $\phi = [\phi(0), \phi(1), \phi(2), \ldots, \phi(N-1)]^{\mathrm{T}}$ and $\beta = [\beta(0), \beta(1), \beta(2), \ldots, \beta(N-1)]^{\mathrm{T}}$ are given by

$$\phi(m) = \frac{1}{f(m) f(N-m) - b(m) b(N-m)}\, f(N-m) \tag{A.16}$$

$$\beta(m) = -\frac{1}{f(m) f(N-m) - b(m) b(N-m)}\, b(m), \tag{A.17}$$

$$m = 0, 1, 2, \ldots, N-1,$$

with $\phi(N) \equiv \phi(0)$ and $\beta(N) \equiv \beta(0)$. In particular, $\phi(0) = 1/f(0)$, $\phi(N/2) = 1/f(N/2)$ and $\beta(0) = \beta(N/2) = 0$.

Proof. To begin with, let us express X explicitly. For $N \geq 4$, using Equation (3.60), X can be graphically depicted as

$$X = \begin{bmatrix} f(0) & & & & & & & \\ & f(1) & & & & & & b(1) \\ & & \ddots & & & & \iddots & \\ & & & f\left(\frac{N}{2}-1\right) & & b\left(\frac{N}{2}-1\right) & & \\ & & & & f\left(\frac{N}{2}\right) & & & \\ & & & b\left(\frac{N}{2}+1\right) & & f\left(\frac{N}{2}+1\right) & & \\ & & \iddots & & & & \ddots & \\ & b(N-1) & & & & & & f(N-1) \end{bmatrix}, \tag{A.18}$$

[1] It is not difficult to prove similar things for odd N but due to the lack of interest (N is usually a power of two) that proof is not shown here.

where blank positions represent zeros. Using the Laplace expansion [29] for the determinant of a matrix along the first row yields

$$\det X = f(0)\ \det \begin{bmatrix} f(1) & & & & & & b(1) \\ & \ddots & & & & \iddots & \\ & & f\left(\frac{N}{2}-1\right) & & b\left(\frac{N}{2}-1\right) & & \\ & & & f\left(\frac{N}{2}\right) & & & \\ & & b\left(\frac{N}{2}+1\right) & & f\left(\frac{N}{2}+1\right) & & \\ & \iddots & & & & \ddots & \\ b(N-1) & & & & & & f(N-1) \end{bmatrix} \tag{A.19}$$

For $N = 4$, a simple application of the Laplace expression again yields the result for the determinant. Otherwise, we proceed. As N is even, the matrix on the right hand side of (A.19) has an odd $(N-1)$ number of lines and columns. If we apply again a Laplace expansion to it in its first row, both $f(1)$ and $b(1)$ terms appear with a plus sign in the expansion:

$$\begin{aligned} \det X = f(0)\Bigg(& \\ & f(1)\det \begin{bmatrix} f(2) & & & & & & b(2) & 0 \\ & \ddots & & & & \iddots & & \vdots \\ & & f\left(\frac{N}{2}-1\right) & & b\left(\frac{N}{2}-1\right) & & & \vdots \\ & & & f\left(\frac{N}{2}\right) & & & & \vdots \\ & & b\left(\frac{N}{2}+1\right) & & f\left(\frac{N}{2}+1\right) & & & \vdots \\ & \iddots & & & & \ddots & & \vdots \\ b(N-2) & & & & & & f(N-2) & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & f(N-1) \end{bmatrix} + \\ & b(1)\det \begin{bmatrix} 0 & f(2) & & & & & & b(2) \\ \vdots & & \ddots & & & & \iddots & \\ \vdots & & & f\left(\frac{N}{2}-1\right) & & b\left(\frac{N}{2}-1\right) & & \\ \vdots & & & & f\left(\frac{N}{2}\right) & & & \\ \vdots & & & b\left(\frac{N}{2}+1\right) & & f\left(\frac{N}{2}+1\right) & & \\ \vdots & & \iddots & & & & \ddots & \\ 0 & b(N-2) & & & & & & f(N-2) \\ b(N-1) & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \Bigg). \end{aligned} \tag{A.20}$$

We will now apply the same expansion to the last row of the matrices in (A.20). These matrices are $(N-2)$-by-$(N-2)$, having thus an even number of rows and columns. Being that the case, the element in the lower right corner appears with a plus sign and the one in the lower left corner with a minus sign. We have thus

$$\det X = f(0)\left[f(1)f(N-1) - b(1)b(N-1)\right] \cdot$$
$$\cdot \det \begin{bmatrix} f(2) & & & & & & b(2) \\ & \ddots & & & & \iddots & \\ & & f\left(\frac{N}{2}-1\right) & & b\left(\frac{N}{2}-1\right) & & \\ & & & f\left(\frac{N}{2}\right) & & & \\ & & b\left(\frac{N}{2}+1\right) & & f\left(\frac{N}{2}+1\right) & & \\ & \iddots & & & & \ddots & \\ b(N-2) & & & & & & f(N-2) \end{bmatrix} . \tag{A.21}$$

For $N = 6$, again a simple application of the Laplace expansion is enough to prove the result. If $N > 6$, we have to apply the Laplace expansion recursively to this matrix the same way we did in equations (A.19) to (A.21), ending up with

$$\det X = f(0) \prod_{m=0}^{\frac{N}{2}-2} \left[f(m)f(N-m) - b(m)b(N-m)\right] \det \begin{bmatrix} f\left(\frac{N}{2}-1\right) & 0 & b\left(\frac{N}{2}-1\right) \\ 0 & f\left(\frac{N}{2}\right) & 0 \\ b\left(\frac{N}{2}+1\right) & 0 & f\left(\frac{N}{2}+1\right) \end{bmatrix} . \tag{A.22}$$

A last application of the Laplace expansion delivers the result (A.14). For $N = 2$, X is of the form

$$X = \begin{bmatrix} f(0) & 0 \\ 0 & f(1) \end{bmatrix} \tag{A.23}$$

and the result is obvious.

To find the inverse of X (if it is invertible) we only have to solve the equation

$$X^{-1}X = \mathrm{I} \tag{A.24}$$

with our candidate X^{-1} for ϕ and β forcing $\beta(0) = \beta(N/2) = 0$ (these two values do not add degrees of freedom to the inverse). If false equations appear, then our candidate is invalid. Nevertheless, as we will see next, no such thing occurs. We get using (A.18) for X and its equivalent for X^{-1} with ϕ and β instead of f and b, respectively,

$$\sum_{k=0}^{N-1}[X^{-1}]_{m,k}[X]_{k,l} = [\mathrm{I}]_{m,l}, \quad m,l = 0,1,\ldots N-1 \iff \tag{A.25}$$

$$\begin{cases} \begin{bmatrix} f(m) & b(N-m) \\ b(m) & f(N-m) \end{bmatrix} \begin{bmatrix} \phi(m) \\ \beta(m) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \text{for } m = 1,2,\ldots,\frac{N}{2}-1,\frac{N}{2}+1,\ldots,N-1 \text{ and} \\ \phi(m)f(m) = 1 & \text{for } m = 0 \text{ or } m = \frac{N}{2}. \end{cases} \tag{A.26}$$

This system has the same number of unknowns and equations and its solution is

$$\begin{bmatrix} \phi(m) \\ \beta(m) \end{bmatrix} = \begin{cases} \begin{bmatrix} f(m) & b(N-m) \\ b(m) & f(N-m) \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \text{for } m = 1,2,\ldots,\frac{N}{2}-1,\frac{N}{2}+1,\ldots,N-1 \text{ and} \\ \begin{bmatrix} \frac{1}{f(m)} \\ 0 \end{bmatrix} & \text{for } m = 0 \text{ or } m = \frac{N}{2} \end{cases} \tag{A.27}$$

$$= \frac{1}{f(m)f(N-m)-b(m)b(N-m)} \begin{bmatrix} f(N-m) \\ -b(m) \end{bmatrix}, \tag{A.28}$$

with $f(N) \equiv f(0)$ and $b(N) \equiv b(0)$.

□