Estimation and Modeling Problems in Parametric Audio Coding

Ph.D. Thesis

MADS GRÆSBØLL CHRISTENSEN

July 2005

Dept. of Communication Technology
Aalborg University
Fredrik Bajers Vej 7
9220 Aalborg Ø, Denmark

Christensen, Mads Græsbøll
Estimation and Modeling Problems in Parametric Audio Coding
ISBN 87-90834-80-1 (print)
ISBN 87-90834-91-7 (electronic)
ISSN 0908-1224

Department of Communication Technology
Aalborg University
Fredrik Bajers Vej 7
DK-9220 Aalborg Ø
Denmark

This thesis was written in LaTeX.

Abstract

The topic of this thesis is parametric coding of speech and audio. A number of estimation and modeling problems in this field of research are addressed. First, a major problem in audio coding, namely efficient coding of transients, is considered. Amplitude modulated sinusoidal models and associated estimators are proposed, and we develop and compare a number of coders. The amplitude modulated sinusoidal models are found to lead to improved coding in listening tests.

Then we move on to the problem of estimating the parameters of sinusoids. We relate a number of practical sinusoidal frequency estimators that are commonly used in audio coding in a framework based on a perceptual distortion measure. These can be related to maximum likelihood estimation under the assumption of Gaussian noise and can be seen as relaxations of the optimal nonlinear least-squares frequency estimator.

The next part concerns a modeling and estimation problem in rate-distortion optimized audio coding. Based on rate-distortion optimization, an optimal segmentation and allocation of bits can be found, but this requires that distortions are calculated for all allocations and segments. We instead propose a method for estimating the distortions based only on a number of simple signal features. Specifically, the relationship between these features and distortions are modeled using a Gaussian mixture, and, for a particular segment, the distortions are estimated using a Bayesian estimator. Listening tests reveal that this can be done without much loss in perceived quality.

Finally, we consider the application of the harmonic sinusoidal model, where all sinusoids are integer multiples of a fundamental frequency, to parametric coding of speech and packet loss concealment based on the sinusoidal parameters. Also, a high-resolution fundamental frequency estimation method is proposed.

The applications of the methods and models presented in this thesis extend beyond the scope of audio coding. The amplitude modulated sinusoidal models and the methods for sinusoidal frequency estimation have applications also in signal and spectral analysis, musical analysis and synthesis, and signal modification.

List of Papers

The main body of this thesis consist of the following papers:

[A] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Amplitude Modulated Sinusoidal Models for Audio Modeling and Coding", in *Knowledge-Based Intelligent Information & Enginerring Systems*, V. Palade, R. J. Howlett, and L. C. Jain, Eds., vol. 2773 of Lecture Notes in Artificial Intelligence, Springer-Verlag, pp. pp. 1334–1342, 2003.

[B] M. G. Christensen, S. van de Par, S. H. Jensen, and S. V. Andersen, "Multiband Amplitude Modulated Sinusoidal Audio Modeling", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, pp. 169–172, 2004.

[C] M. G. Christensen and S. van de Par, "Efficient Parametric Coding of Transients", to appear in *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(4), July 2006.

[D] M. G. Christensen and S. H. Jensen, "Computationally Efficient Amplitude Modulated Sinusoidal Audio Coding using Frequency-Domain Linear Prediction", submitted to *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2006.

[E] M. G. Christensen, A. Jakobsson, S. V. Andersen, and S. H. Jensen, "Linear AM Decomposition for Sinusoidal Audio Coding", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, pp. 165–168, 2005.

[F] M. G. Christensen and and S. H. Jensen, "On Perceptual Distortion Minimization and Nonlinear Least-Squares Frequency Estimation", to appear in *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(1), pp. 99–109, January 2006.

[G] F. Nordén, M. G. Christensen, and S. H. Jensen, "Open Loop Rate-Distortion Optimized Audio Coding", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3., pp. 161–164, 2005.

[H] C. A. Rødbro, M. G. Christensen, F. Nordén, and S. H. Jensen, "Low Complexity Rate-Distortion Optimized Time-Segmentation for Audio Coding", in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 231–234, 2005.

[I] C. A. Rødbro, M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Compressed Domain Packet Loss Concealment of Sinusoidally Coded Speech", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 104–107, 2003.

[J] M. G. Christensen, S. H. Jensen, S. V. Andersen, and A. Jakobsson, "Subspace-based Fundamental Frequency Estimation", in *Proc. European Signal Processing Conf.*, pp. 637–640, 2004.

The following papers and patents have been published and filed, respectively, by the author of this thesis during the Ph.D. studies:

[1] M. G. Christensen, C. Albøge, S. H. Jensen, and C. A. Rødbro, "A Harmonic Exponential Sinusoidal Speech Coder", in *Proc. NORSIG-2002, 5th Nordic Signal Processing Symposium*, 2002.

[2] M. G. Christensen and S. van de Par, "Rate-Distortion Efficient Amplitude Modulated Sinusoidal Audio Coding", in *Conf. Rec. Thiry-Eighth Asilomar Conference on Signals, Systems, and Computers*, pp. 2280–2284, 2004.

[3] M. G. Christensen, A. Jakobsson, S. V. Andersen, and S. H. Jensen, "Amplitude Modulated Sinusoidal Signal Decomposition for Audio Coding", accepted for publication in IEEE Signal Processing Letters.

[4] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Improving Sinusoidal Modeling Using Squared Instantaneous Envelope for Obtaining Amplitude Shape", Danish patent application number 2003 01231, applicant: Aalborg University, filed August 2003.

[5] M. G. Christensen and S. van de Par, "Parametric Audio Coding Comprising Amplitude Envelopes", European patent application number 04105428.9, applicant: Koninklijke Philips Electronics N.V., filed November 2004.

Maxims of Signal Processing

Julius O. Smith III, professor at Stanford University, formulated the following humorous fundamental principles of signal processing.

1. Everything is equivalent to everything else, once you finally understand it.
2. If one technique is superior to another, it is due to a longer integration time (more averaging).
3. Exciting new results are usually due to artifacts in the processing.
4. With enough processing, it is no longer necessary to have any input data.
5. Scale factors are never right on the first try.

Preface

This thesis is submitted to the International Doctoral School of Technology and Science at Aalborg University in partial fulfillment of the requirements for the degree of doctor of philosophy. The main body consists of a number of papers that have been published in or been submitted to peer-reviewed conferences or journals. The work was carried out during the period August 2002–July 2005 at the Department of Communication Technology at Aalborg University, and it was funded by the ARDOR (Adaptive Rate-Distortion Optimized sound coderR) project, EU grant no. IST-2001-34095.

It is customary that, in the preface of a Ph.D. thesis, the author thanks everybody he ever met. I will here seek to limit myself to those who made a significant contribution to my work and life over the past three years, and still then many people deserve to be acknowledged. First of all my supervisor Søren Holdt Jensen deserves my gratitude for giving me this opportunity, believing in me, and for giving me the freedom to pursue my interests and do what I do best. Also, I am grateful to my co-supervisor Søren Vang Andersen who contributed greatly to this work through our many discussions and by encouraging me to explore the amplitude modulated sinusoidal audio coding. I would also like to point out that the work of my "intellectual ancestor" here at Aalborg University, Jesper Jensen, is what got me started working on the amplitude modulated sinusoidal modeling and coding, and for that he also deserves to be recognized.

This thesis is to a large extent the result of collaboration with other people, and my various co-authors thusly also deserve honorable mention here. First of all, Steven van de Par deserves to be thanked for our fruitful collaboration and my stay at Philips Research Labs in Eindhoven, The Netherlands. I thank Andreas Jakobsson for spurring my interest in estimation theory and our collaboration on various topics and papers. Besides the papers we wrote together, I also thank Christoffer A. Rødbro for the many discussions both technical and otherwise, and Fredrik Nordén for our collaboration that unfortunately was cut short.

I would also like to recognize my present and former "brothers in arms" here at Aalborg University Karsten V. Sørensen, Xuefeng Yin, Joachim Dahl, Morten H. Larsen, Steffen Præstholm, Troels Pedersen, Chunjian Li, and Bin Hu for their significant technical and/or social contributions to my life these past three years.

I thank the people at KTH, France Telecom, Hannover University, Delft Technical

University, and Philips who were involved in the ARDOR project for the countless interesting technical discussions and the fun times we had at our numerous meetings. I also thank Tonny Gregersen for the effort he put in the C implementation of the amplitude modulated sinusoidal audio coder

Last but not least, I thank my friends and family, but most of all Majbritt for love and support the past three years.

Mads Græsbøll Christensen
Aalborg, July 2005

Contents

Introduction

1 Source Coding

1.1 Introduction

Compression of audio signals, be it music or speech, can be described as the art of achieving the highest possible perceived quality of audio signals represented using a given number of bits or, conversely, minimizing the number of bits required to encode a signal at a given quality. Despite the availability and decreasing cost of high rate transmission channels, the interest in perceptual audio coding remains high today. The reason is simple. Instead of making perceptual audio coding obsolete, increasing bandwidth instead opens up new applications and possibilities. For example, many companies have seized the opportunity to spawn a variety of new consumer electronics products such as portable digital audio players that are much smaller, lighter and more flexible than previous products based on compact discs or cassette tapes. Digital audio broadcasting (DAB) [1] and streaming over the Internet are also becoming more popular and important in today's society. Many high quality audio systems now rely on multiple channels, not just stereo, and this is yet another example of the new possibilities. It also seems logical that since channel capacity will always be limited and never without cost, we will continue to seek to get the most from a given channel, and, hence, the source coding problem persists. For an overview of some of the many applications of audio coding, we refer to the papers [1–7]. It is also interesting to note that while early audio and video players and recorders were mainly based on a single codec, there now seems to be a development towards support for many different codecs. This opens the door for great progress since the research community then no longer necessarily has to concern itself with backward compatibility.

Audio coding belongs to the field of source coding, which is part of communication and information theory. First, we treat the audio coding problem as a source coding problem. Later, we will present the specifics of audio coders and go into details about parametric coding of audio. Finally, the contributions of this thesis will be presented.

1.2 The Shannon Communication Model

It was C. E. Shannon who, with his landmark 1948 papers [8, 9], laid the foundation of modern communication theory. A block diagram of the Shannon model of communication is shown in Figure 1. The source produces messages, in our case discrete-time audio signals, that are to be transmitted to the sink. In Shannon's work, the source is characterized by a probability density function (pdf) or a probability mass function (pmf). The source is transformed into a different signal, a string of bits, by the source encoder. The source encoder seeks to find a representation that exploits statistical properties of the source such that the number of bits required is minimized. This string of bits is then passed on to the channel encoder. The function of the channel encoder is to protect the signal from corruption by the transmission channel. This is done by adding redundant information that allows for error correction in the channel decoder. The source and channel encoders are jointly referred to as the encoder. The signal is then transmitted through the communication channel or is stored on a medium.

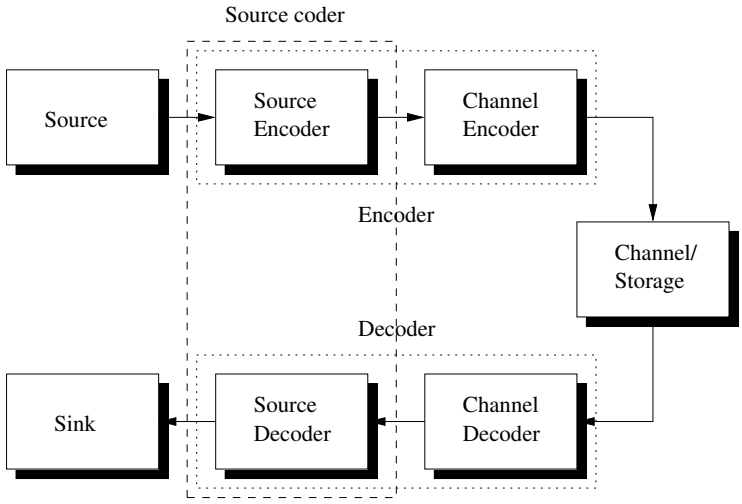

Figure 1: The Shannon Communication Model.

The transmission channel is the physical medium that the signal is transmitted through on its way to the sink. This may, for example, be the acoustical channel between two people communicating in a room or a radio frequency channel. The channel may subject the transmitted signal to noise. In the channel decoder, the redundant information that was added by the channel encoder is exploited to correct errors caused by the noisy channel and thereby recovering the original signal. This is then fed to the source decoder which maps the signal back to a form similar to that of the original

source. It then finally arrives at the sink. The combination of the source and channel decoders are commonly referred to as the decoder while the combination of a encoder and a decoder is commonly known as a codec or a coder.

One of the main results of Shannon's work is that the source and channel coding processes can be separated. This, however, is only strictly true for infinite delay and complexity, and as a result, joint source and channel encoding may be of interest in some applications. For example, the nature of the transmission channel is taken into account in recent work on speech coding for packet based networks [10–15]. For the most part, in fact except for paper I, this thesis is mostly concerned with source encoding and decoding, commonly referred to as source coding. For a survey of the results and research activities in source coding, we refer to [16–18].

1.3 Design Criteria

The design of a coder is typically subject to a number of explicit or implicit constraints that depends on the application. Some design criteria are conflicting and their relative importance depends on the application. These must all be taken into account in the design or evaluation of a coder (see e.g. [4, 19]).

Rate: The number of bits per second. The rate may either be fixed or variable over time. Often coders are designed to operate at specific rates, and which coder is the best depends on the desired bit-rate. For example, parametric coders are known to perform better than transform coders at low bit-rates while at high bit-rates, transform coders perform best (see e.g. [20, 21]).

Distortion: The quality of a coder is measured in terms of distortion. The higher the distortion, the worse the quality. Distortion may be measured objectively, but ultimately the subjective quality is what matters. This is determined using listening tests.

Delay: The delay is the time it takes from the message is produced at the source until it arrives at the sink. The higher the delay, the lower the rate is achievable at the same distortion. The delay constraints on the encoder/decoder may vary greatly depending on the application. For example, in music storage applications, no hard delay constraint exists and the encoder may take advantage of this, while for telephone systems there are strict requirements for the tolerable delay.

Memory: The amount of memory that is required by the encoder/decoder. How much memory is required in the encoder and/or decoder is also an important factor in many applications. Often memory can be traded for complexity and vice versa.

Complexity: The number of CPU/DSP cycles required by the encoder/decoder to process a block of data. Many applications of audio coding require real-time encoding and/or decoding of the coded signal and often the decoder has to run on devices that

have very little processing power such as mobile phones or portable players. Hence, the computational complexity is often considered critical for the decoder.

Robustness: The ability of the coder to handle errors and anomalies in the input. Some transmission channels are subject to bit errors while others are subject to losses of entire packets (i.e. the frame erasure channel). Hence, the encoder/decoder should be designed with robustness towards the respective types of errors. Robustness is traded for bit-rate.

Flexibility: How well the coder adjusts to different constraints and/or conditions. For example, rate scalability is desirable in many applications. The ability of a coder to adjust to different constraints and input signals is becoming increasingly important. For example, the available bandwidth on the Internet may be different at different times of the day, or the probability of packet losses may vary. Speech coders can code speech very efficiently at very low bit-rates but they do not perform well for music. Audio coders can code both music and speech well, but at higher bit-rates. The price paid for flexibility is often complexity and/or bit-rate.

The art of source coding is then to arrive at a reasonable tradeoff between all these factors for the application at hand.

1.4 Fundamentals

Let us start by introducing the basics of source coding. Since all practical coders operate under the constraint of finite delay, we here concern ourselves with signals of finite length, namely blocks of signals. It is also assumed that we are dealing with discrete-time signals, i.e. signals that have been sampled in an appropriate manner. The input signal is denoted as $x(n)$ and the vector containing a block of the input signal as $\mathbf{x} = [\, x(0) \;\ldots\; x(N-1) \,]^T \in \mathbb{R}^N$. A vector quantizer (VQ) is then a mapping $Q : \mathbb{R}^N \rightarrow \mathcal{C}$ of $\mathbf{x}$ to a codebook (CB) $\mathcal{C}$, i.e. we write $Q(\mathbf{x}) = \hat{\mathbf{x}}_i$. The case where $N = 1$ is referred to as scalar quantization. The codebook consist of a number of reproduction vectors $\mathcal{C} = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots \hat{\mathbf{x}}_K\}$ with $\hat{\mathbf{x}}_i \in \mathbb{R}^N$ for $i \in \mathcal{I} = \{1, 2, \ldots, K\}$. The aim of vector quantizer design is then for a given codebook size K (or entropy) to design a codebook that minimizes the error that the quantization process causes. No optimal solution exists for this problem, however, and a wealth of different algorithms for codebook design exists (see e.g. [22, 23]). With respect to the Shannon communication model, the vector quantizer can be described in terms of an encoder that maps the input vector to a member of the index set, i.e. $E : \mathbb{R}^N \rightarrow \mathcal{I}$. This index maps to a bit string that is transmitted. The decoder is then the map $D : \mathcal{I} \rightarrow \mathbb{R}^N$ back from the bit string corresponding to a member of the index set to the reconstruction vector. From these definitions, it can be seen that any coding system can be classified as a vector quantizer. In fact, this is also the argument used to show that vector quantization is indeed optimal [17, 22, 24]. The differences between different quantizers or coders

are then in terms of the process that maps from the input signal to the index and back again. A vector quantizer is (at least locally) optimal (see [22]) if its encoder, given a codebook, operates in such a way that it satisfies the *nearest neighbor condition* meaning that a vector is mapped to the nearest reconstruction vector in the sense of some distance (distortion) measure and the codebook satisfies the *centroid condition*, meaning that the reconstruction vectors are the ones minimizing the expected distortions of the partition cells that define the decision boundaries between neighboring reconstruction vectors. Finally, the *zero probability boundary condition* must be satisfied, that is, boundary points (exactly equally distant to two reconstruction vectors) occur with zero probability.

1.5 Definition of the Coding Problem

Let us now move on in defining the source coding problem mathematically. The problem at hand can be described as finding a reconstruction vector $\hat{\mathbf{x}}_i$ of the input signal $\mathbf{x}$. In posing the source coding problem, a nonnegative distortion measure $D(\mathbf{x}, \hat{\mathbf{x}}_i)$ that measures how close the reconstruction vector is to the input signal is needed. The distortion measure must satisfy a number of properties to be a metric or norm (see [25]). Source coders can be described as either distortion-constrained or rate-constrained. Distortion-constrained means that the source coding problem can be posed as

$$\begin{aligned} \text{minimize} \quad & R(i) \\ \text{s.t.} \quad & D(\mathbf{x}, \hat{\mathbf{x}}_i) \leq D^\star, \end{aligned} \tag{1}$$

where $D^\star$ is the desired distortion. In this setup, the number of bits used is minimized while the distortion is equal to or less than some required level. For rate-constrained coding, the mathematical optimization problem is

$$\begin{aligned} \text{minimize} \quad & D(\mathbf{x}, \hat{\mathbf{x}}_i) \\ \text{s.t.} \quad & R(i) \leq R^\star, \end{aligned} \tag{2}$$

with $R^\star$ being the desired number of bits. Here, the distortion is minimized while the number of bits is at most $R^\star$. Both of these constrained optimization problems can be solved using the Lagrange multiplier method (see e.g. [18, 26, 27]) and this is commonly referred to as rate-distortion optimization. How this is done in practice is described in several of the papers in this thesis.

The rate $R(i)$ associated with transmitting the codebook indices can be measured in terms of resolution (codebook size) or entropy. When measured in terms of the resolution (number of different representations), the resulting rate is upper bounded and fixed, whereas when measured in terms of entropy, the resulting expected rate will be lower and variable. The entropy of a source is defined as the average minimum number of bits required to reconstruct it without any loss [8, 9]. Specifically, the entropy of a

random variable I with (quantizer indices) outcomes $i \in \mathcal{I}$ is defined as

$$H(I) = -\sum_{i \in \mathcal{I}} p(i) \log_2 p(i), \tag{3}$$

with the individual indices having probability $p(i)$. Furthermore, there exist entropy codes of expected length L such that $H(I) \leq L \leq H(I) + 1$ (see [18]). Hence, we can then design vector quantizers subject to an entropy constraint rather than a resolution constraint and deal with the entropy coding in a separate step. Entropy coding is also commonly referred to as lossless coding. Lossless coding is a type of distortion-constrained coding where $D^\star = 0$ and where $D(\mathbf{x}, \hat{\mathbf{x}}_i) = 0$ implies that $\hat{\mathbf{x}}_i = \mathbf{x}$. Here, the goal is to minimize the rate under the condition that the input signal can be reconstructed perfectly. Common examples of lossless coding are Huffman, arithmetic and Ziv-Lempel coding (see e.g. [18, 22]). Parametric and perceptual audio coders belong to the class of lossy coders where $D^\star \geq 0$, which in turn implies that we may have that $\hat{\mathbf{x}}_i \neq \mathbf{x}$. Lossy coders often also apply lossless coding in the final stage of the encoding process. Using these definitions, we can also define perceptually transparent coding as distortion-constrained coding with $D^\star = 0$, but where we may have that $D(\mathbf{x}, \hat{\mathbf{x}}_i) = 0$ for $\hat{\mathbf{x}}_i \neq \mathbf{x}$. An alternative definition of transparent coding, with $D(\mathbf{x}, \hat{\mathbf{x}}_i) = 0$ if and only if $\hat{\mathbf{x}}_i = \mathbf{x}$, would be $D^\star = \epsilon$, where the distortion measure is then constructed in such a way that a distortion less than ϵ is guaranteed to be inaudible.

In practice, most lossy source coders are rate-constrained. The formulation in (2) of the source coding problem comes naturally from transmission channels or storage media having limited capacity. Given the constraints of the transmission channel or storage medium, we seek to find the best encoding of a given signal. In this case, best is measured in terms of the distortion $D(\cdot)$. Basically, we seek to do the best with what we have.

In designing source coders, we then face a number of tasks. We have to either choose or design 1) a distortion measure, 2) codebook vectors, and 3) an encoder and decoder. The function of the codebook is to describe the signal in a compact way whereas the function of the distortion measure is to shape the error such that it has the least impact on the perceived quality.

1.6 Rate-Distortion Theory

Shannon derived bounds on the possible minimum rate required to encode a source at a given distortion [28]. We will here briefly state the main results. For proofs and details we refer to [17, 18, 24, 28]. Here, we consider $x \in \mathcal{B}$ to be the discrete outcomes of a random variable X having a pmf $p(x)$. The quantization process then results in another random variable $\hat{X} = Q(X) \in \mathcal{C}$. The expected distortion of the quantization process

can then be written as

$$E\left\{D(X,\hat{X})\right\} = \sum_{\hat{x}\in\mathcal{C}}\sum_{x\in\mathcal{B}} p(x,\hat{x})D(x,\hat{x}) \tag{4}$$

$$= \sum_{\hat{x}\in\mathcal{C}}\sum_{x\in\mathcal{B}} p(x)p(\hat{x}|x)D(x,\hat{x}). \tag{5}$$

The entropy of the reconstruction point or vector $\hat{X}$ can be related to the mutual information between $\hat{X}$ and X as

$$H(\hat{X}) \geq H(\hat{X}) - H(\hat{X}|X) \tag{6}$$

$$= I(X;\hat{X}), \tag{7}$$

where the conditional entropy is zero, i.e. $H(\hat{X}|X) = 0$, since we are not concerned with noisy channels or reconstruction. This means that the entropy of the quantized vector, or, equivalently, the entropy of the quantization indices, simply equals the mutual information. The mutual information between X and $\hat{X}$ is defined as

$$I(X;\hat{X}) = \sum_{\hat{x}\in\mathcal{C}}\sum_{x\in\mathcal{B}} p(x,\hat{x})\log_2\frac{p(x|\hat{x})}{p(x)}. \tag{8}$$

Since $p(x)$ is given, the problem of minimizing the entropy associated with $\hat{X}$ reduces to the design of a statistical mapping, a quantizer, having the conditional probability $p(\hat{x}|x)$. Indeed, rate-distortion theory states that if we pick this such that

$$R(D) = \min_{p(\hat{x}|x):E\left\{D(X,\hat{X})\right\}\leq D} I(X;\hat{X}) \tag{9}$$

we can do no better. This minimum rate as a function of the distortion is known as the rate-distortion function. The rate-distortion function is a convex, non-increasing function of the distortion. The rate-distortion bound can be attained by using long (in principle infinite) vectors. That is, for sufficiently long segments, there exists codes such that the rate-distortion bound is attained. It turns out that even when quantizing independent random variables a lower rate can be achieved by vector quantization than by quantizing the variables independently [18]. Given these results, we can justifiably claim that all practical source coding, other than unstructured quantization of arbitrarily long vectors, is motivated by implicit delay and complexity constraints.

The Shannon rate-distortion theory [28] is often considered nonconstructive (see e.g. [22]) since it does not give us a practical method for designing vector quantizers. For real-life signals, the pdf of the source is generally not known.

1.7 Parametric Coding

The topic of this thesis is parametric audio coding. The basic principle of parametric coding is essentially an extrapolation of Occam's razor that one should not use any

more parameters to describe a system/signal than necessary (see, e.g., [18]). Or, in other words, the simplest possible model that can be used to describe the system/signal is probably also the right one.

We weakly define parametric coding to mean compression by means of modeling a signal with a few physically and/or psychoacoustically meaningful parameters $\boldsymbol{\theta} \in \mathbb{R}^L$ (see also [29]). The number of parameters L is less than the length of the input signal vector N and in most cases much less, i.e., we have that $L \ll N$. The parametric encoder is then a map from the input signal vector to the (intermediate) parameters, and through quantization of these parameters, to the index which is transmitted. The parametric decoder is the map back to the parameters and to the reconstruction vector. By a few parameters, we mean that the number of parameters are less, and often much less, than the number of input samples. One distinct feature that separates parametric coding from vector quantization using trained codebooks is that the signal model $\hat{x}(n, \boldsymbol{\theta})$ is *chosen* (or postulated, some may say) by the designer. From the point of view of vector quantization, one might say that parametric coding is a form of structured vector quantization where the codebook structure is imposed by the signal model. The basic sentiment shared among people working on parametric coding seems then to be that we can better choose a signal model than design a codebook training algorithm. That the codebook design is reduced to a choice of signal model can generally be attributed to difficulties in imposing codebook structures on the training algorithms. Likewise, that the model parameters are chosen such that the quantization of model parameters maps to well-studied experiments in psychoacoustics is due to shortcomings in the distortion measures. In reality, though, the success and popularity of parametric coding is due to a number of factors. Namely, we can choose models such that the parameters can be found in a computationally efficient way, and, for the case of audio coding, we can choose models that are expressed in terms of parameters whose sensitivity to quantization errors are well-studied in the psychoacoustical literature. Mathematically, the parametric coding problem can be defined as follows: Let the input signal $x(n)$ be segmented as $\mathbf{x} = [\, x(0) \, \cdots \, x(N-1) \,]^T \in \mathbb{R}^N$. Then we seek to find a description of that input signal $\hat{\mathbf{x}}(\boldsymbol{\theta}) = [\, \hat{x}(0, \boldsymbol{\theta}) \, \cdots \, \hat{x}(N-1, \boldsymbol{\theta}) \,]^T \in \mathbb{R}^N$ in terms of some parameters $\boldsymbol{\theta} \in \mathbb{R}^L$.

A parametric coder typically consists of a number of processing components, namely model parameter estimation, perceptual modeling that calculates a distortion measure, parameter quantization, and entropy coding. These are depicted in the block diagram in Figure 2.

It is interesting to note that our definition of parametric coding is contradictory to the popular distinction of audio coders into parametric and waveform coders. That distinction is clearly misleading since parametric coding can be, and often is, waveform approximating (e.g. [30–32]). However, most non-waveform approximating coding techniques are parametric. Waveform approximating coders can have the useful property that the reconstructed signal converges to the original as the bit-rate grows. As a result, they generally require higher bit-rates than the non-waveform approximating

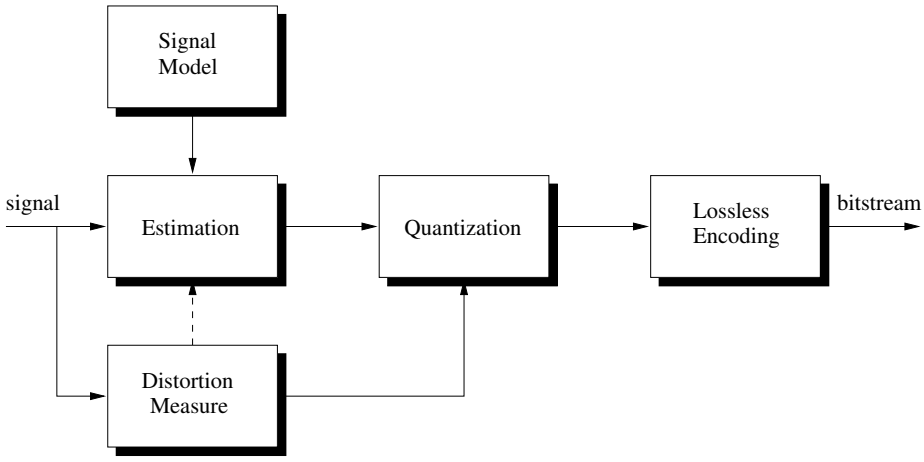

Figure 2: Parametric encoder. Given a signal model, the parameters are estimated from the observed signal. The model parameters are then quantized and entropy coded before they are written to the bitstream. A perceptual model is used to calculate a distortion measure, which can then be applied in the estimation and quantization of model parameters.

coders, but they are then typically also more robust, i.e. they can handle a wide variety of different sources [21]. Furthermore, it can be seen that, by our definition of parametric coding, it can generally not achieve perfect reconstruction since $L < N$. However, many models used in parametric coding, for example the sinusoidal model, can include $L = N$ as a special case. Most coding standards and practical coders today employ combinations of different coding techniques and many also use parametric coding methods.

1.8 Estimation and Modeling Problems

Having defined the source coding problem and parametric audio coding, we can now define what is meant by the title of this thesis, namely estimation and modeling problems in parametric audio coding. A modeling problem can be defined as the problem of finding a signal model, which is suitable for some purpose, or, in terms of vector quantization, designing a codebook having a certain structure. The purpose considered here is signal compression and we hence seek a model that results in efficient coding of audio signals. Estimation problems, on the other hand, are concerned with, given a signal model, finding the model parameters as well as possible. The measure of goodness of the model and estimated parameters is that the distortion $D(\mathbf{x}, \hat{\mathbf{x}}(\boldsymbol{\theta}))$ is minimized while the number of bits associated with the parameters $\boldsymbol{\theta}$ is kept fixed. Often, the quantization of model parameters or compressed-domain signal modifications, such as [33],

rely on the physical interpretation of the parameters. In that case, the statistical properties of the estimated parameters may also be of interest. Here, it should be noted that when a number of coders are used in a multi-stage structure, such as [21, 34], where the output of one coder is subtracted from the input and this difference is fed to a different coder, one cannot simply conclude that the individual stages should, greedily, minimize the distortion. Hence, the estimation-theoretical perspective of finding the best parameters (see e.g. [35, 36]), in a statistical sense, may also be applicable to parametric coding.

2 Audio Coding

2.1 Brief History

Audio coding is one of the great success stories of modern digital signal processing. Most people in the Western world today have heard about, or use, "mp3" (MPEG-1 Layer III [37]) frequently and, despite its age, it remains the de facto audio coding standard. Although speech and audio coding had been an active research field for many years before that, e.g. [38], the major breakthrough in audio coding came in 1990 when J. D. Johnston and K. Brandenburg made a historic demonstration at AT&T Bell Labs. They demonstrated the concept of perceptually transparent coding of audio by spectral noise shaping. They compared two audio signals with the same signal-to-noise ratio (SNR), namely 13.6 dB. For one of the signals, the noise was white, while for the other signal, the noise was shaped, in the frequency domain, by the auditory masking threshold [39]. A SNR of 13.6 dB is typically considered as very low quality. Indeed, most analog-to-digital converters used in audio processing system employ 16 of 24 bits per sample, corresponding to SNRs of 96 or 144 dB. The difference in perceived quality between the two signals was such that the demonstration has since been dubbed "The 13 dB miracle". Perceptual audio coding is often described as the task of removing/exploiting redundancies and irrelevancies of audio signals for compression purposes. The redundancies refer to the statistical properties of the source, e.g. statistical dependencies over time, while the second refers to taking the properties of the human auditory system into account. Most people today accept the compact disc (CD) quality as the benchmark against which compression schemes should be compared. The CD uses 16 bits per sample with a sampling frequency of 44.1 kHz and the resulting rate is 705.6 kbps for a mono signal. Even by today's standards, streaming or processing data at such rates would be impossible, or at least very expensive, in many applications. Both lossless and lossy coding of audio have been considered throughout the years. It turns out, however, that only moderate savings in bit-rate, a factor of two or three, can be achieved with lossless coding [40]. As a consequence, many applications require that lossy coding be applied.

For a complete survey of the different standards and methods for perceptual audio

coding, we refer to the tutorials [20, 41, 42] and for specifics of speech coding, we refer to the tutorials [43, 44] and the references therein.

2.2 Perceptual Noise Shaping

In this section, we illustrate the basic principles of perceptual audio coding and the notions of perceptual distortion measures and noise shaping. We do this based on the simplest possible quantizer, a uniform scalar quantizer. First, the signal is transformed into a perceptual domain by the, possibly nonlinear, transformation $T(\cdot)$. In that domain, the signal is quantized and mapped back to the signal domain using the inverse transformation $T^{-1}(\cdot)$. The transformation may depend on the signal $x(n)$ as is often the case in audio coding. This is illustrated in a block diagram in Figure 3. Noise shaping is also sometimes constructed such that $T^{-1}(T(x(n))) = x(n)$, in which case the original can be reconstructed perfectly.

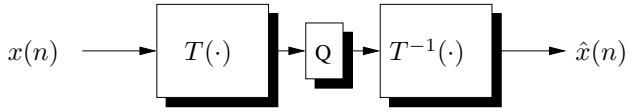

Figure 3: Noise shaping by quantization in the perceptual domain.

The quantization error caused by the quantization process (denoted Q) can be modeled as an additive noise process $e(n)$. In the block diagram in Figure 3, this means that the quantization block is replaced by an addition of the noise process $e(n)$. This is depicted in Figure 4. For a uniform scalar quantizer, the noise process $e(n)$ is white and has a uniform probability density function (pdf); the number of bits determines the variance of $e(n)$. Now, the role of the transformation $T^{-1}(\cdot)$ also becomes evident. It will shape the error according to the transformation $T^{-1}(\cdot)$; so, for the same quantizer, the error can be shaped in many different ways. Clearly, it is desirable that the error is shaped such that it has least impact; in the case of audio signals, this means that the error is introduced where it is the least audible. This is the basic idea of noise shaping, and it is very fundamental to audio coding. At this place, it is important to realize that the signal-to-noise ratio cannot be improved by noise shaping. Noise shaping is often also used in audio analog-to-digital converters and in wordlength reduction, but usually in much simpler ways than in audio coders, e.g. [45–49].

From Figures 3 and 4, it can also be seen that we could just as well have designed the quantizer such that the error was shaped according to $T^{-1}(\cdot)$ rather than transforming the input and output of the quantizer. This would, however, mean a more complicated quantizer design.

The public switched telephone network (PSTN) also employs the principle of noise

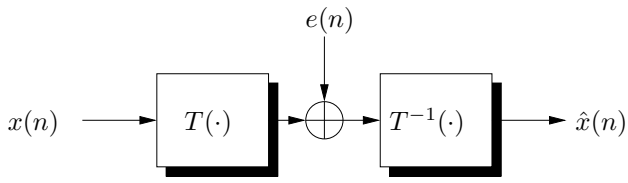

Figure 4: Noise shaping by transformation of the input and output.

shaping, but in a very simple way. In the 1972 ITU-T standard for audio and speech companding G.711 (see e.g. [43, 50]), a logarithmic transform is applied to the input, which is subsequently quantized uniformly. Some, though, e.g. [22], see this as motived by the source pdf not being uniform. The pre- and postfiltering approaches of [51, 52] also fit this description well. In [51, 52] the transformation $T(\cdot)$ is a linear filter that implements a spectral weighting and in [53, 54], the error is shaped in the temporal domain in order to improve coding of transients.

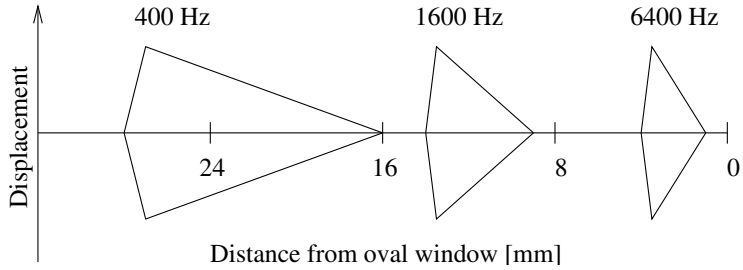

Figure 5: Illustration of frequency-to-place transformation along the basilar membrane for a tone complex with the curves indicating the envelopes of the traveling waves (crudely after [55]).

2.3 The Human Auditory System

What is special about perceptual audio coders is that they take the properties of the human auditory system into account in the coding of signals [4, 56–58]. The human auditory system is a complex and highly nonlinear system. The ear consists of three parts: the outer, middle and inner ear. The function of the outer ear is to collect the signal and pass it on to the middle ear that serves as a transducer. Here, the acoustical waves are transformed into compressional waves in the fluid in the inner ear. In the inner ear these are transformed into nerve impulses that are transmitted to the brain. Audio coders take advantage of the effects, or limitations, of the processing in the human auditory system.

Transparent coding aims at shaping the error in such a way that it is inaudible. The simplest possible way of doing this is by taking advantage of the absolute threshold of hearing, that is the minimum level of a tone that is audible. This level depends greatly on the frequency of the tone. Any spectral components below the absolute threshold of hearing are not audible and hence do not have to be coded. Audio coders also take advantage of a phenomenon known as masking. Masking can be defined as a reduced audibility of a signal known as the maskee due to the presence of another signal known as the masker. In terms of coding this means that spectro-temporal perturbations (quantization or modeling errors) that may be inaudible can be introduced. The maximum inaudible perturbation is known as the masking threshold. The notion of masking threshold led to the definition of perceptual entropy in [4, 59, 60], which is the bit-rate required to encode a signal in such a way that it is indistinguishable from the original. However, fairly complicated signal analysis is required in order to derive the masking thresholds and the concept of masking is not applied easily to all types of audio coders. As an example of a model for calculating such masking thresholds, we refer the reader to the ISO 11172-3 (MPEG-1) Psychoacoustic Model 1 [37] also described in [20].

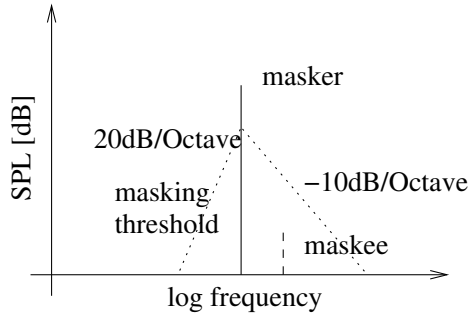

Figure 6: Illustration of simultaneous masking with the sound pressure level (SPL) as a function of log-frequency. The presence of a strong tone, the masker, make a nearby tone, the maskee, inaudible if it is bellow the auditory masking threshold.

The literature on psychoacoustics (see e.g. [55, 61]) distinguishes between two types of masking: spectral (or simultaneous) and temporal (nonsimultaneous) masking. Simultaneous masking is the steady-state masking that occurs in the presence of a stationary masker. Simultaneous masking can be understood by observing the frequency-to-place transformation that takes place along the basilar membrane in the cochlea of the human ear. This is illustrated for a tone complex consisting of three tones in Figure 5. It can be seen that the different tones affect different locations of the basilar membrane and that the effect is not well localized. This part of the human auditory system can be seen as filter bank consisting of a set of tuned filters that adapt to the stimuli. What can be observed is that the energy in one region spreads to neigh-

boring filters, and because of this spread a nearby less strong signal may not be audible. This is known as the spread of masking and its properties are illustrated in Figure 6. The figure illustrates that, in the presence of a sinusoid having high amplitude, another sinusoid, with a smaller amplitude at a nearby frequency, can become inaudible and hence does not have to be coded. There are, however, some problems in applying the concept of masking to audio coding since the derived masking thresholds depend on the type of masker and the type of maskee. For example, the sensitivities encountered in experiments with tones masking noise versus noise masking noise are different. Also, the human auditory system is highly nonlinear. This can, for example, be observed in that the response to a certain stimulus depends on the loudness level. This also leads to the problem that, at the time of the encoding, it is not known what the playback level will be. Hence, we see that the masking analysis inherently will be subject to a number of choices and tradeoffs.

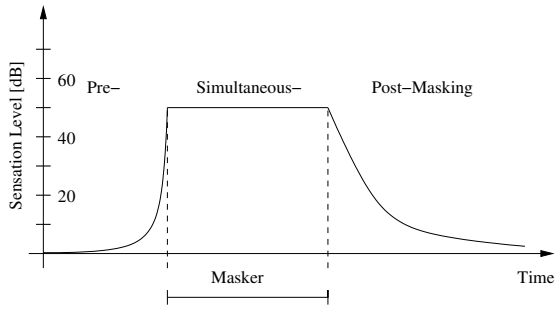

Figure 7: Illustration of nonsimultaneous masking (crudely after [55]). In the presence of the masker, simultaneous masking occurs while before and after the masker, nonsimultaneous masking occurs.

Nonsimultaneous masking refers to the phenomenon that masking effects can be observed before and after a masker is present. The masking capabilities of the masker build up as the sensitivity decreases. Right before the onset of the masker, the sensitivity decreases rather quickly. After the masker has ended, the sensitivity recovers rather slowly. This is known as pre and post-masking, respectively and is depicted in Figure 7. The post-masking is a much stronger effect than pre-masking, and, therefore, modeling or quantization noise introduced before the onset of a signal component can be especially troublesome. These kinds of errors are usually referred to as pre-echos or pre-echo distortion and much effort has been put into solving this problem (see e.g. [20, 53]), including papers A through E in this thesis.

In non-transparent coding, e.g. rate-constrained coding, it is not a matter of what is masked and what is not. Rather, it is a matter of shaping the error in a way such that the error is least audible and this leads naturally to the need for perceptually motivated distortion measures.

2.4 Generalized Linear Distortion Measure

The masking properties of the human auditory system have been discussed, but how these concepts can be applied in a distortion measure has yet to be addressed. We now proceed to define the distortion measure $D(\mathbf{x}, \hat{\mathbf{x}}(\boldsymbol{\theta}))$ in a fairly general but mathematically convenient way. We define it as a generalized 2-norm, which can be written as

$$D(\mathbf{x}, \hat{\mathbf{x}}(\boldsymbol{\theta})) = \|\mathbf{W}(\mathbf{x} - \hat{\mathbf{x}}(\boldsymbol{\theta}))\|_2^2, \tag{10}$$

with $\mathbf{W} \in \mathbb{R}^{M \times N}$ being a perceptual weighting matrix, which may depend on $\mathbf{x}$ and $\hat{\mathbf{x}}(\boldsymbol{\theta})$. Using this formulation, all of the R-D optimization problems turn out to be some form of least-squares. Writing out the 2-norm, we get

$$D(\mathbf{x}, \hat{\mathbf{x}}(\boldsymbol{\theta})) = (\mathbf{x} - \hat{\mathbf{x}}(\boldsymbol{\theta}))^H \mathbf{W}^H \mathbf{W} (\mathbf{x} - \hat{\mathbf{x}}(\boldsymbol{\theta})), \tag{11}$$

where $\mathbf{W}^H\mathbf{W}$ is now guaranteed by construction to be symmetric and positive semidefinite. We can now write the eigenvalue decomposition of this as [62]

$$\mathbf{W}^H\mathbf{W} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H, \tag{12}$$

with $\mathbf{U} \in \mathbb{R}^{N \times N}$ containing the column eigenvectors and $\boldsymbol{\Lambda} \in \mathbb{R}^{N \times N}$ the associated (nonnegative) eigenvalues λ_i on the diagonal. If we have that all the eigenvalues are positive $\lambda_i > 0 \; \forall i$, the distortion measure defines a norm, whereas if any of the eigenvalues are zero, i.e. $\lambda_i = 0$, we have a pseudo-norm. In that case, errors in the subspace of the orthogonal complement of the row space of $\mathbf{W}$ will have zero distortion. The perceptually weighted two-norm can now be written in the following form

$$D(\mathbf{x}, \hat{\mathbf{x}}(\boldsymbol{\theta})) = (\mathbf{x} - \hat{\mathbf{x}}(\boldsymbol{\theta}))^H \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H (\mathbf{x} - \hat{\mathbf{x}}(\boldsymbol{\theta})). \tag{13}$$

From this it is also clear that in our choice of signal model $\hat{\mathbf{x}}(\boldsymbol{\theta})$ it is desirable that the matrix-vector product $\mathbf{U}^H\hat{\mathbf{x}}(\boldsymbol{\theta})$ can be calculated efficiently. We have still not discussed how to derive the perceptual weighting matrix $\mathbf{W}$. This is an active field in current psychoacoustical research. Much effort has been put into defining perceptually meaningful distortion measures. For example, the model presented in [63, 64] defines such a measure, where the perceptual weighting matrix $\mathbf{W}$ is circulant and symmetric. For such a matrix, the eigenvectors are the well-known Fourier basis and such a structure is a desirable property in that it admits efficient computations. In [65, 66], a fairly complicated and nonlinear signal processing model of the human auditory system is presented. It was originally designed to predict the outcome of masking experiments, but in [67] this measure is linearized and a perceptual weighting matrix is derived from it. This matrix does not have the circulant structure of the model in [63], and, hence, the calculation of distortions is more complicated. Other examples of perceptually motivated distortion measures and auditory models for audio coding are given in [68–70]. The model of [65, 66] is monaural but auditory models for the binaural case have also

been investigated, for example [71]. Also in the field of speech coding the importance of the distortion measure has been recognized [72–74]. How such a distortion measure can be incorporated in sinusoidal audio modeling and coding in an efficient way is the topic of paper F.

2.5 Transform/Subband Coding

Transform coding is based on a linear map from the signal domain to another domain by means of an invertible transform, often a unitary transform. In the transform domain, the signal is then quantized and mapped back to the signal domain in the decoder by the inverse transform. The function of the transform is in terms of energy compaction and decorrelation. It is desirable to choose a transform such that the energy is now concentrated in only a few parameters, whereby low-dimensional vector quantizers or even scalar quantizers may be applied. Under a Gaussian assumption, the optimal transform can be shown to be the Karhunen-Loeve transform (see [22]). However, the Karhunen-Loeve transform is signal dependent and thus not very useful for coding purposes as the transform would need to be transmitted to the decoder. Generally, due to the nonstationary nature of audio signals, no one transform is optimal for all segments of a given signal. For speech and audio signals, different variations of the sinusoidal transforms have been used, such as the Fourier transform and the discrete cosine/sine transforms; although wavelets have also been considered in e.g. [75, 76]. Transform/subband coding is by far the most successful coding technique for audio coding (e.g. [77–80]) and such coders have been standardized in the context of MPEG (Moving Picture Experts Group) [37, 41, 56, 81–83]. In Figures 8 and 9, block diagrams of generic audio encoder and decoders are shown, respectively.

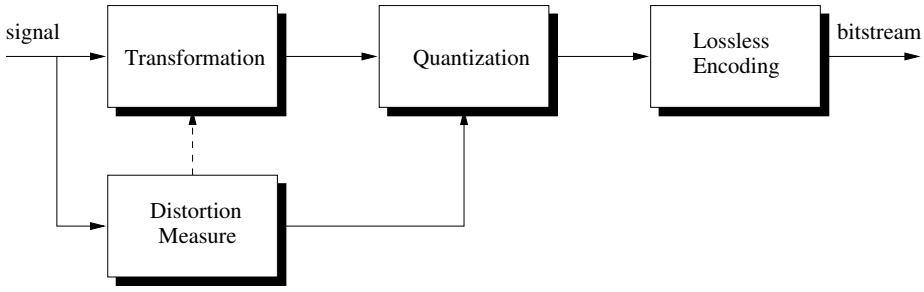

Figure 8: Generic audio encoder. The signal is first transformed into a set of model or transform parameters or coefficients. These parameters are quantized according to a perceptual distortion measure which is formed by an analysis of the input signal using a perceptual model. The quantized parameters are entropy coded and written to the bit-stream.

Most state-of-the-art transform audio coders are based on the modified discrete cosine transform (MDCT), for example the MPEG-2/4 AAC (Advanced Audio Cod-

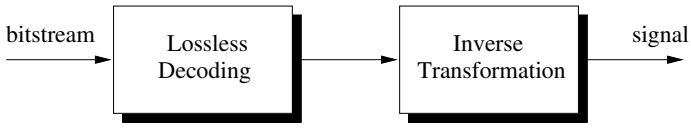

Figure 9: Generic audio decoder. The quantized parameters are reconstructed from the bitstream by the entropy decoder and the signal is reconstructed from these parameters by the inverse transformation.

ing) [82–84]. The MDCT is a type-IV discrete cosine transform (DCT) [85]. It is a critically sampled lapped transform based on the principle of time-domain aliasing cancellation [86, 87]. The MDCT (and many other similar transforms) can also be interpreted as perfect reconstruction modulated filter banks. Therefore, transform coding can also be seen as a form of subband coding. The MPEG-4 AAC algorithm is generally considered state-of-the-art and forms the measure against which new coders are measured in MPEG. It exists in a number different incarnations (known as profiles): main, low complexity, scalable sampling rate, long term prediction, low delay and high efficiency. The AAC scheme incorporates a number of different coding techniques and supports a wide range of different bit-rates.

2.6 Quality Assessment

The quality of audio and speech coders is primarily evaluated in listening tests, where a number of subjects are asked to grade processed, i.e. coded, excerpts on some scale relative to the unprocessed excerpt. The primary reason for this is that ,even though much effort has been invested in research in objective measures e.g. [88–92], these still fail to predict the outcome of listening tests under different conditions. In listening tests, it is imperative that appropriate test conditions and methods be defined such that the results can be reliably compared. This is exactly what the recommendations in e.g. [93, 94] do. For example, in [93], methods for screening (and rejection) of listeners are given. An often used set of excerpts for evaluation of audio coders are those on the EBU SQAM (Sound Quality Assessment Material) discs [95]. Often expert listeners are preferred over inexperienced listeners in assessing the quality of an audio coder. The reason for this is that over time, inexperienced listeners will, with extensive exposure to a certain coding technique, tend to become more sensitive to coding artifacts. Hence, experts listeners give an accurate indication as to whether a coder will stand the test of time [96]. A natural question seems also to be how to define transparent quality. Transparent quality means that the test samples are indistinguishable from the originals. Specifically, the term was defined by the EBU as "*A decoded test item is indistinguishable from the reference test item when, using the triple stimulus hidden reference method, the 95 % confidence intervals of the test and reference subjective assessments overlap*" [97].

In the papers in this thesis, three types of listening tests have been used. The tests

have been carried out as proof of concept and are not to be considered formal. In determining whether one, supposedly improved, system outperforms another system, we use a preference test where the listeners are asked which of two processed excerpts they prefer relative to a reference without knowing which is which. Typically, especially when dealing with small differences, such tests are repeated. It is important that such tests are randomized in terms of the order of presentations and the excerpts. In determining the overall quality of different coding techniques MUSHRA-like tests have been employed [93]. In these kind of tests, the listeners are asked to rank a number of processed excerpts on a scale from 0 to 100 relative to the original with a score of 100 indicating that the processed excerpt is identical, in quality, to the original. Also, a number of additional excerpts processed in a well-defined way, known as anchors, and a hidden reference are included. For evaluation of the packet loss concealment methods for sinusoidally coded speech considered in paper I, listening tests were also carried out based on a five point degradation score [94], where the listener is asked to rate the degradation due to different channel conditions relative to the reference.

3 Parametric Coding

3.1 Introduction

Although many coding techniques presently used in combination with transform coders fit the defintion of parametric coders, we here primarily focus on a specific type of parametric coding, namely sinusoidal coding. Sinusoidal coding is also often combined with other types of coding, such as residual (or noise) coding. Historically, the interest in sinusoidal coding started during the search for efficient coding of speech signals for mobile telephony. In the 1980s, much research was devoted to the application of sinusoidal modeling of speech for coding purposes [33, 98–105] and somewhat later also to its applications to musical analysis and synthesis [106–108]. Since the mid 1990s there has been much interest in standardization bodies such as MPEG (Moving Picture Experts Group) and research in parametric coding [109–119]. The sinusoidal model has also been applied to speech enhancement [120] and more recently, renewed interest in sinusoidal coding of speech has been spurred by the increasing interest in voice over packet-based networks [12–15, 121], where sinusoidal coders form an attractive alternative in that frame independence can more easily be achieved compared to LPC-based coders [43].

The most general form of sinusoidal coding is based on the following signal model, where a segment is modeled for $n = 0, \ldots, N-1$ as

$$\hat{x}(n) = \sum_{l=1}^{L} a_l(n) \cos(\Phi_l(n)) + w(n), \tag{14}$$

where the signal is modeled as a deterministic part, which is composed of a sum of

sinusoids, and a stochastic part $w(n)$ [29, 108]. The sum of L sinusoids each characterized by an instantaneous amplitude $a_l(n)$, also known as the amplitude modulating signal in modulation theory, and an instantaneous phase $\Phi_l(n)$. The stochastic part is often modeled as an auto-regressive moving-average (ARMA) process

$$w(n) = \sum_{i=0}^{I} b_i e(n-i) + \sum_{f=1}^{F} c_f w(n-f), \tag{15}$$

with $e(n)$ being the excitation signal, often modeled as Gaussian noise having a possibly time-varying variance, and b_i being the MA coefficients and c_f the AR coefficients.

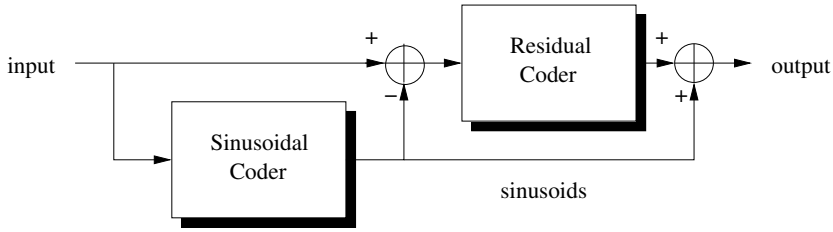

Figure 10: Typical parametric coder consisting of a sinusoidal and residual coder.

Typically, the decomposition into a deterministic and a stochastic component is done in a sequential way, where first all sinusoids are extracted and parameters quantized in sinusoidal coder and then the remaining signal, also known as the residual, is fed to a so-called noise or residual coder that codes the stochastic parts. This is illustrated in Figure 10. Each of the sinusoidal and the residual coding blocks can be described as in Figure 8. This kind of coding is sometimes also referred to as multi-stage coding. Transient signal segments are often also handled by a separate coder in a multi-stage (or switched) structure, e.g. [114–116, 122]. Some parametric coders are only based on the deterministic part [33, 99], but we will return to that later.

3.2 Sinusoidal Models

The model stated in (14) is quite general and not in itself all that useful for coding purposes. In this section we take a look at sinusoidal model variations that have been applied in parametric coding, and some that have only been applied to modeling. Here, the reader must bear in mind that just because a model can capture the energy of a signal, it does not mean that it is also efficient in terms of bit-rate.

We have not yet put any restriction on the phase $\Phi_l(n)$ and the amplitude $a_l(n)$ of (14). The frequency of the sinusoid is defined as the derivative of the instantaneous phase, i.e. $\omega_l(n) = \frac{\partial \Phi_l(n)}{\partial n}$. Many different models of the amplitude and the phase have

been proposed. The simplest and perhaps most successful is where both the frequency and the amplitude are constant for a particular segment, i.e.,

$$\hat{s}(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l), \tag{16}$$

where ϕ_l is the starting phase of the l'th sinusoid. This model is very well suited for modeling of stationary tonal signals such as voiced speech, trumpet, violin and many other signals. For periodic signals the relation between the frequencies of the individual components may be modeled as $\omega_l = l\omega_0$ where ω_0 is the fundamental frequency. This model, which is referred to as the harmonic sinusoidal model, has been widely used in speech modeling and coding, e.g. [98, 123–125]. In some applications it is of interest to relax this constraint somewhat. For example, in [126–128], it was noted that higher quality modeling and coding of speech can be achieved using a quasi-harmonic model where $\omega_l = l\omega_0 + \Delta_l$ with Δ_l being a small deviation for the l'th component from the integer multiples of the fundamental frequency. It is also well-known that certain instruments, for example stiff-stringed instruments such as the piano, produce harmonics that are not exact integer multiples of a fundamental frequency [129], and for this reason, the quasi-harmonic model has also been applied to modeling of audio signals [106, 130].

The harmonic sinusoidal model is too restrictive for general audio and as a consequence the unconstrained model where ω_l may take on any value is typically preferred. Also, different phase representations have been considered. For voiced speech, an onset model of the phase, where the phase of the individual components are assumed to be synchronous in the residual domain of a linear prediction (LP) filter, has been reported to produce reasonable results [98, 131, 132]. In this model, the instantaneous phase is modeled as $\Phi_l(n) = \omega_l(n - n_0) + \phi_l$ with n_0 being the onset and ϕ_l then the minimum-phase contribution of the LP filter. Perhaps the most efficient coding of phases arise from sinusoidal components evolving slowly from segment to segment. Then the phase of a sinusoidal component can be predicted from the previous segment by taking the frequency change into account. This is known as phase prediction and has been widely used in sinusoidal coding of speech [98, 133–137]. The caveat is that it introduces strong interframe dependencies and is hence sensitive to packet losses and bit errors. The minimum-phase assumption of LP has also been known to cause problems in this context.

It is well-known that the speech production system is not always minimum-phase, for example for nasal sounds [138–140]. As a result of this, phase compensation by means of an all-pass filter has also been investigated and applied to sinusoidal coding of speech [140, 141].

Many different models have been proposed for efficient coding of transients, though very few have actually been proven to improve on existing coders. Different variations of damped sinusoids where the instantaneous amplitude (or amplitude modulating sig-

nal) can be expressed as $a_l(n) = u(n - n_l)A_l \exp(-\beta n)$ where $u(n)$ is the unit step function, have been studied [123, 142–147], mainly because of their mathematically convenient form. Some of these methods, namely [142, 143, 147], impose the constraint that $n_l = 0$ for all l such that the components may only start at the beginning of a segment. The model proposed in [145, 146] allow any n_l, but the resulting envelopes are not smooth. That smoothness generally is desirable is the reason that overlap or interpolative synthesis is used to avoid discontinuities between segment boundaries in all audio coders, also when optimal segmentation is used [148].

A different aspect of this is that the steepness of the attack is also an important factor in the recognition of musical instruments [149]. In the sinusoidal coder described in [21, 117–119], which is the current reference parametric coder in MPEG [150], this is recognized as the amplitude modulating signal is modeled using so-called Meixner functions.

Some methods for audio modeling do not impose any specific model on the amplitude modulating signal, but rather assume that it is slowly varying [105, 106], but these are not directly applicable to coding. A rather different method that can also be seen as an amplitude modulating signal model was proposed in [53], although for noise shaping purposes. As we shall see in paper D, this can also be applied to amplitude modulated sinusoidal audio coding. Different realizations of low-order polynomial phase and amplitude modeling have also been considered [29, 151–153] and these have been reported to improve on the perceived quality especially for nonstationary voiced speech.

3.3 Sinusoidal Parameter Estimation

Given the wealth of different sinusoidal models, it is not surprising that also a great number of different estimators exist; in some cases, the model has perhaps been chosen because an efficient estimator exists. Indeed, the freedom in choice of models and estimators has made this area an active area for signal processing researchers. The problem of estimating the parameters of a set of sinusoids in noise arises in many different applications, and is therefore also well-studied and well-documented.

For the sinusoidal model, consisting of a sum of sinusoids with each having a constant phase, a constant amplitude and a frequency for a particular segment, all of these parameters have to be estimated. It turns out, however, that the amplitude and phase parameters can be found in a straigh-forward manner using linear least-squares [154]. The difficult part is the nonlinear part, namely the frequencies, and for some models also the parameters of the modulating signal. One can distinguish between a number of types of frequency estimators: a) subspace-based methods b) nonlinear least-squares (NLS) methods c) filter bank methods and d) rational spectra methods. Only a) and b) have been widely used in parametric audio coding and modeling; we will discuss only those approaches. For an overview of the other methods, and frequency estimation in general, we refer to the excellent book on spectral estimation [36].

In subspace-based methods, the eigenvalue decomposition of the sample covariance

matrix is decomposed into a noise subspace and a signal (plus noise) subspace (see e.g. [36, 155, 156]). The various subspace methods either model the structure of the signal subspace or exploit that all sinusoids belonging to the signal subspace are orthogonal to the noise subspace. Among the subspace-based methods, MUSIC [157–159], ESPRIT [160] and the subspace fitting method(s) [156] are especially noteworthy. ESPRIT can also be used for estimating damped sinusoids.

The NLS class of estimators are based on the principle of least-squares, meaning that they seek to minimize a squared error measure, generally to achieve consistency with respect to the source coding problem formulation. This is by far the most widely used principle in parametric coding. The optimal estimator is then a multidimensional numerical search for the L sinusoids that minimize the squared error, but such an exhaustive search simply not feasible due to the computational complexity. This optimal estimator is known as the NLS estimator [36, 161, 162]. Due to the intractable complexity of the NLS method, most of the practical methods find sinusoids iteratively, one at a time. Matching pursuit [85] is perhaps the most famous example of this. For more information on matching pursuit and its various derivatives, we refer to [29]. It is interesting to note that the nonlinear least-squares method also can be seen as a subspace pursuit, where the target subspace is spanned by L sinusoids [163]. Also, the often used, and much maligned, peak picking method [33, 99], which is based on the discrete Fourier transform, can be seen as a least-squares method under certain conditions.

Estimators are traditionally evaluated in terms of the bias and variance of the estimates (see e.g. [35, 36]). In the case of a model mismatch the estimates will generally be biased. A lower bound of the variance of unbiased estimators, known as the Cramér-Rao bound (CRB), can be found given the properties of the noise. Assuming that the noise is white and distributed according to a Gaussian pdf, it can be shown that minimization of a squared error measure is the same as estimating the most likely parameters, i.e. least-squares estimation is equivalent to maximum likelihood estimation. When the noise is Gaussian but colored, the maximum likelihood estimator corresponds to a weighted least-squares estimator. However, for the sinusoidal signal model, least-squares and maximum likelihood estimation is asymptotically (for many observations) the same [36, 161, 162].

There has recently been much interest in incorporating perception into these estimators. Audio modeling and coding have in common that it is of interest to find a compact representation, or in other words to represent the signal in as few, physically meaningful parameters as possible. Since the receiver of these signals is the human auditory system, it is also of interest to represent the perceptually most important components. In audio coding in particular, it is desirable to estimate and transmit only the parameters of audible sinusoids; in recent years, much effort has been put into this problem, e.g. [32, 143, 147, 164–169]. Often, these methods rely on more or less heuristic rules taken from psychoacoustic experiments. In [166], for example, sinusoidal components are found in an iterative manner by assigning a perceptual weight to the spectrum and then picking the most dominant peak of the weighted spectrum. Another method is

the so-called pre-filtering method [51, 52, 143], where the observed signal is filtered using a perceptual filter in order to achieve a weighting of the sinusoidal components. The methods of [167] and [168] are different methods yet that rely on loudness and excitation pattern similarity criteria for sinusoidal component selection, respectively.

In coding applications, it is of particular interest to state the estimation criterion in a way that defines a distortion measure or metric. A globally optimal solution that minimizes this distortion measure ensures that at a given bit-rate (for a certain number of sinusoids in the case of sinusoidal coding), the lowest possible distortion is achieved. When the distortion measure is a perceptual one, meaning that it reflects the behavior of the human auditory system, we can then claim that the perceived distortion is minimized at the given bit-rate. The psychoacoustic (or perceptual) matching pursuit [165] is a descendant of matching pursuit algorithm where the 2-norm is replaced by a perceptually motivated distortion measure [63, 64]. Rather than converging in the 2-norm, the algorithm converges in a perceptually more meaningful way. Then, we also see that there are some parallels to the analysis-by-synthesis methods using perceptually motivated distortion measures often used in speech coding, e.g. [170]. Matching pursuit can be seen as an analysis-by-synthesis method with the special requirement that the distortion measure is induced by an inner product.

3.4 Sinusoidal Parameter Quantization

The quantization of model parameters has also been the subject of much research. Many of the problems involve incorporating psychoacoustics into the quantization process. As a result, many coders, for example [148], employ quantizers that are derived from just-noticeable-differences (JNDs) in psychoacoustical experiments (see e.g. [55, 61]) and these are thus not based on a proper distortion measure that may be subject to optimization. The consequence is that these kinds of quantizers lead to inflexible coders. The JNDs are also typically found in experimental setups that are very simple compared to the application.

Recently, there has been some interest in the application of high-rate quantization theory (see e.g. [22]) to sinusoidal coding [34, 171–177]. The high-rate assumption leads to analytical expressions for the optimal point densities of a quantizer for a given target-rate. This is an extremely flexible approach that allows many degrees of freedom in the optimization of the distribution of bits over components and segments compared to fixed quantizers. Also, the so-called spherical or polar quantizers of [34, 171–173, 176, 177] incorporate joint quantization of the sinusoidal parameters such that, for example, the number of bits used in quantizing the phase of a component depends on its amplitude and perceptual weight. Vector quantization of sinusoidal parameters has also been considered, but there is a problem in that the length of the vector is not always the same. In fact, it may vary substantially. In [178], this is handled by a variable-to-fixed length transformation for the amplitudes. Differential encoding of the sinusoidal parameters has also been of interest. The phase prediction mentioned earlier,

perhaps combined with encoding of the prediction error, is in fact such a method. The problem of optimal time or frequency differential encoding of the sinusoidal parameters has been extensively studied and solutions are compared and proposed in [179–181].

3.5 Residual Coding

A sum of a small number of sinusoids is not an effective model for all signals. Indeed, noise-like stochastic signals are not well-modeled using sinusoids; the alternate coders used to compress such signals are known as residual or noise coders. In speech coding, though, there has been a tradition of modeling the stochastic part using sinusoids as well (e.g. [33, 98, 99]), mainly because it is convenient. Actually, any signal, including broadband noise and transients, can be modeled using sinusoids that are spaced "close enough", but it is just not efficient in terms of bit-rate, and any spectral subsampling will inherently result in temporal aliasing. In practice, acceptable quality can be achieved using this approach for narrowband speech, although it often results in the synthesized speech being very tonal. For wideband speech, unvoiced parts are not well modeled using sinusoids, and fricatives sound especially bad.

In audio coding, different encoding methods must be used for handling noise-like signals. Typically, the encoding of the deterministic and stochastic components is handled in a multi-stage structure where first a number of sinusoids are extracted and subtracted from the input. The remaining signal, called the residual, is then encoded by a different coder.

Residual coders can be grouped into those that are waveform approximating and those that are not. The first group generally operates at higher bit-rates than the last group. Examples of waveform-approximating residual coders are [174, 182] where a sinusoidal coder is combined with an MDCT-based transform coder, and in [21], the residual is coded using regular-pulse excitation (RPE) [183]. It is interesting to note that in [21] the parametric coder is demonstrated to outperform MPEG4-AAC at 64 kbps (stereo). Residual coders belonging to the other group, i.e. those that do not encode the signal in a waveform-approximating way, are sometimes called noise coders. They are based on the notion that the human auditory system cannot distinguish between two realizations of the same noise process. Typically the spectral and temporal envelopes are encoded. The noise coders generally rely on either an ARMA model, or derivatives thereof, [107, 108, 111, 117–119, 184, 185] or perceptually motivated filter bank-/transform-based synthesis [29, 114, 122, 186–189]. Perceptual noise substitution [190, 191] is also such a method. Some work on incorporating perception into the ARMA model-based coders has also been published [184, 192–194].

The waveform approximating coders have the desirable property that the reconstructed signal may converge to the original signal as the bit-rate is increased. As a result, the waveform approximating coders generally perform better at higher bit-rates while the so-called noise coders perform best at low bit-rates. The problem of quantization of AR parameters is also well-studied in the context of speech cod-

ing [74, 195, 196].

3.6 Applications of Rate-Distortion Optimization

Rate-distortion optimization is a valuable tool for source coding. Here we will focus mainly on the application of rate-distortion optimization to parametric coding, although it has also been applied to transform coding in e.g. [197, 198]. By rate-distortion optimized coding, we mean coding that is optimized according to the input signal and rate or distortion constraints. This is based on the so-called operational rate-distortion theory [199] and is based on the results of [26] that the optimization problem of rate-constrained coding can be solved using the Lagrange multiplier method even for discrete allocation problems (see e.g. [200]). Operational rate-distortion theory is different from the Shannon rate-distortion theory [28] in that the outcomes of the stochastic processes that produce the signals are the subject of optimization rather than statistical properties such as the expected distortion.

Rate-distortion optimized coding parts with arbitrarily chosen bit allocations and tradeoffs. Instead, these problems are solved using constrained optimization. For example, it is possible to distribute bits optimally over segments, provided that the delay constraint allows it, such that more efficient coding is achieved. In [27, 199] an algorithm for segmenting an input signal in a rate-distortion optimal way was devised and applied to linear predictive coding of speech signals. Traditionally, audio and speech coders use either a fixed segment length or find a variable segmentation based on heuristic stationarity measures. The window-switching method [201] used in e.g. [202, 203] is an example of this. Such methods can neither adapt to different signals nor coders.

In the last few years, these principles have been applied to parametric audio coding [148, 175, 204, 205]; and are likewise used in the papers in this thesis. One major drawback of the optimal segmentation is that it requires the encoding of all different segment sizes at all possible starting points. This is extremely wasteful in terms of computational complexity since the final encoding only will make use of a small subset. There has also been some recent interest in rate-distortion optimized multi-stage coding of audio, e.g. [182]. Rate-distortion optimized coding is sometimes considered not to be feasible due to delay or complexity constraints. In that case, it may be considered a benchmark, a development tool, for investigating and even quantifying the loss in performance due to different constraints.

3.7 Relation to Vector Quantization

By the definition of a vector quantizer, we see that a sinusoidal coder too can be seen as a vector quantizer of sorts. This interpretation is further strengthened by the similarity between a particular method for vector quantization and a method for sinusoidal coding. Matching pursuit [85] is an algorithm for iteratively building a signal model. It has a wide variety of applications including model-based signal compression and signal

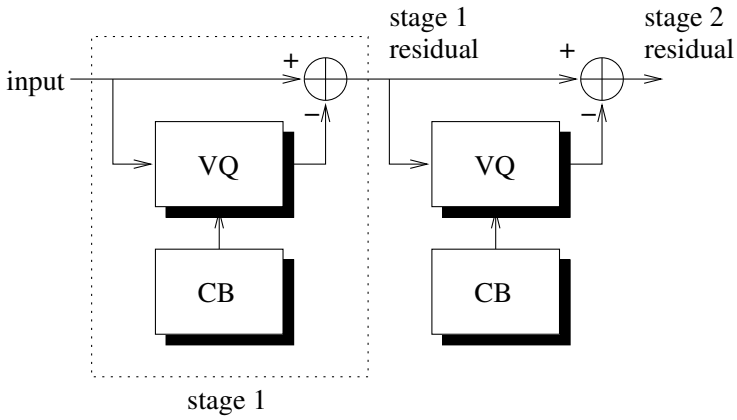

Figure 11: Multi-stage vector quantizer (here two stages). The reconstruction vector of each stage is subtracted from the input and the difference is fed to the next stage.

analysis, and it has been widely used for sinusoidal modeling and coding of audio [29, 144, 163, 165, 166, 204, 205]. In each iteration, the matching pursuit selects an element (a vector) from a dictionary (a codebook) that with an optimal scaling best matches the input signal in the sense of some norm. This scaled dictionary element is then subtracted from the input and the process is repeated. In sinusoidal modeling and coding, the dictionary is built from complex sinusoids having different frequencies and the associated scalings are then the amplitude and phase of the sinusoids.

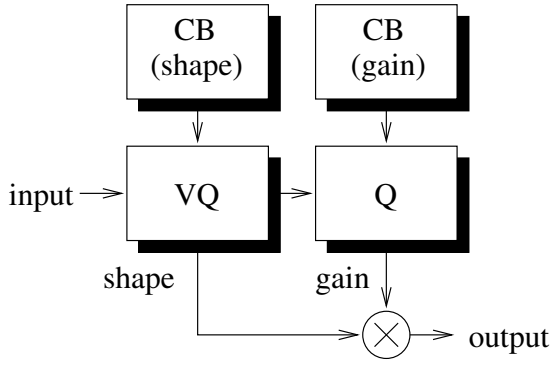

Figure 12: Shape-gain vector quantizer. The reconstruction vector is a scaled codebook shape.

Multi-stage quantizers are quantizers consisting of multiple stages where in each

stage, a vector quantization operation is performed; the codebooks may or may not be the same in the different stages. The reconstruction vector of a stage is then subtracted from its input and the resulting signal, called a residual, is then fed to the next stage. This is illustrated in Figure 11. A shape-gain vector quantizer is a vector quantization method where a reconstruction vector is found along with an optimal scaling. In this way, the energy of the reconstruction vector is treated separately from the shape (as if the pdf of the shape is independent of the gain), whereby the complexity involved in the VQ can be minimized significantly. The structure of a shape-gain vector quantization encoder is shown in Figure 12. It is readily seen then that by constructing a quantizer from multiple stages of shape-gain vector quantizers, we get a structure similar to that of matching pursuit. The difference is basically one of terminology.

From the discussion above, it should be clear that a sinusoidal coder using matching pursuit with subsequent quantization of the phase and amplitude can be seen as a multi-stage shape-gain vector quantizer, where the shape index corresponds to a sinusoid while the gain corresponds to the amplitude and phase. The codebook is a highly structured one constructed from complex sinusoids of different frequencies. From the point of view of vector quantization, perhaps the dictionary should not be chosen but rather designed using codebook training algorithms. This is indeed the idea in [206, 207]. Then, however, it would by our definition no longer belong to the class of parametric coders.

3.8 Other Parametric Coders

Besides the incarnation of parametric coding that we have just discussed, i.e. the combination of a sinusoidal coder and a residual coder, there are other types of coding techniques that can be characterized as parametric. In fact, many are already in use in standardized audio coders such as MPEG-4 AAC. For example, spectral band replication (SBR) [208] and perceptual noise substitution (PNS) [190, 191] are coding techniques that may be characterized as parametric. Also, parametric representations are also often used in multi-channel coding (see e.g. [209]). For example, binaural cue coding (BCC) [210, 211] is a technique where the stereo channels are derived parametrically from a mono channel as opposed to the often used non-parametric sum-difference coding [203]. Other examples of parametric stereo coding are the so-called intensity stereo coding [212, 213] and the more recent extension to BCC [214, 215]. There have also been activities in e.g. MPEG on extending these principles to more than two channels [216] and these efforts continue today [209]. Early linear predictive speech coders, such as LPC-10 [217], were often also described in a parametric fashion, where the signal is separated into a vocal tract contribution and a parametrically modeled excitation signal [43]. However, this interpretation may be considered somewhat of a stretch for waveform approximating coders such as [170].

4 Contributions

The title of this thesis is quite broad because the topics and contributions are rather diverse, but all apply to parametric audio coding and modeling. Papers A through E deal with amplitude modulated sinusoidal audio coding, a coding technique for efficient coding of transients. Papers C, E, G, and H are about, or use, coding based on rate-distortion optimization using a perceptual distortion measure, and paper F deals with sinusoidal frequency estimation using this perceptual distortion measure. Paper I is concerned with the application of the harmonic sinusoidal model for speech coding for packet based networks and how packet loss concealment can be achieved using the sinusoidal parameters; Paper J deals with the estimation of the fundamental frequency of the harmonic sinusoidal model. We will now go through the contributions of the individual papers that constitute the main body of this thesis.

Paper A: This paper introduces the notion of amplitude modulated sinusoidal audio modeling from the perspective of modulation theory. The paper is based on Bedrosian's separation theorem of carriers and modulating signals [218] using the Hilbert transform and the analytic signal (see e.g. [219]) and treats different signal manipulations and models in this framework. One of the strengths of this framework is that it is fairly general with respect to the assumptions that are made regarding the signal.

Paper B: In this paper, the theory of paper A is brought to practice in an audio modeling system. This system is used to study the importance of allowing different amplitude modulating signals for different frequency regions; it is confirmed in listening tests and using a model of the human auditory system [65, 66] that significant improvements are achieved by this. The paper does not, however, deal with the question of whether this is also efficient in terms of bit-rate.

Paper C: An amplitude modulated sinusoidal audio coder is developed. It is based on a model of the modulating signal which is characterized by an onset, an attack, and a decay. Each sinusoidal component can have a different envelope. This model is combined with a sinusoidal coder without amplitude modulation in a rate-distortion optimized framework that uses optimal distribution of sinusoids over segments and optimal segmentation [26, 27]. The amplitude modulated sinusoidal coder is shown to improve on a baseline coder in listening tests. This work proves that it is indeed efficient in terms of bit-rate to allow different modulating signals for different components and that optimal segmentation and adapted models are complementary coding techniques; furthermore, the optimal segmentation changes with the signal model.

Paper D: In this work, we develop an amplitude modulated sinusoidal audio coder based on the theory of Paper A and the results of paper B. We use frequency-domain linear prediction, a principle similar to the temporal noise shaping of [53], as a means for estimation and efficient coding of the modulating signal. This coder has very low complexity and requires little memory compared to that in Paper C, and it is

demonstrated to improve upon a baseline coder in a delay constrained setup. Like in paper A, the strength of the methods used in this paper is that the model of the modulating signal is not very restrictive.

Paper E: In this paper, the amplitude modulating signal is modeled as a linear combination of arbitrary basis vectors. This model is rather different from those of papers B-D in that the constraints on the amplitude modulating signal being nonnegative is relaxed; sinusoidal frequencies may occur at spectral minima. The model can exploit spectral symmetries for coding purposes and is demonstrated in listening tests to improve upon a sinusoidal coder. Also, this coder has the advantage that since the model parameters are linear they may easily be optimized.

Paper F: Here, we develop a framework for frequency estimation based on a perceptual distortion measure [63, 64]. We relate a number of different practical estimators [51, 143, 166] in this framework and investigate how they relate to estimation theory. Seen in the light of this work, the pre-filtering method for incorporating perception in estimators and the weighted matching pursuit can be seen as approximations to the optimal perceptual nonlinear least-squares method (see e.g. [36, 161]) and can be shown to be equal to the perceptual, or psychoacoustic, matching pursuit [165] under certain conditions.

Paper G: In this paper, we apply the framework of [220] to sinusoidal coding. The basic idea is to estimate the distortion of a sinusoidal audio coder at different rates given a set of observed features. The dependence between the features and the distortion is modeled in a probabilistic setting where the joint probability density is modeled using Gaussian Mixtures and the distortions are estimated using a Bayesian estimator.

Paper H: We extend the work of paper G further in this paper by addressing a number of problems in the framework for estimating the distortion at different rates based on signal features. We apply the framework to the problem of finding the rate-distortion optimal segmentation. This results in a significant complexity reduction in finding the optimal segmentation and it is demonstrated, via listening tests, that this can be done without much loss of performance.

Paper I: Recently, there has been much interest in speech coding for packet based networks where packets may be lost. We here develop a parametric coder specifically for speech based on the harmonic sinusoidal model, wherein all sinusoids have frequencies that are integer multiple of a fundamental frequency. We perform packet loss concealment based on time-scale modification using the sinusoidal components. A simple listening test shows that graceful degradation as a function of the packet loss probability can be achieved.

Paper J: In this paper, we present a method for estimation of the fundamental frequency of the harmonic sinusoidal model used in paper I. The method is based on an

eigenvalue decomposition of the sample covariance matrix and a subsequent separation into signal and noise subspaces. We propose a method where the rank, and hence the number of harmonics, of the subspaces depend on the fundamental frequency, and the estimate is found as the fundamental frequency for which the harmonically related sinusoids are the closest to being orthogonal to the noise subspace. An interesting observation here is that the fundamental frequency can, generally, be estimated more accurately than the frequency of any of the individual harmonics.

It seems natural to ask, based on the contributions of this thesis, what general conclusions can be drawn and in what direction audio coding will be heading in the future. We have shown that it is possible to significantly improve parametric coders by dedicated signal models, such as modulated sinusoids, for certain critical signals. It is the opinion of the author that the results of this thesis prove that rate-distortion optimized coding can succeed and that the concept of perceptual distortion measures can be applied successfully in audio coding. Furthermore, the author believes that continued research in perceptual distortion measures is the key to future advances. It has also been demonstrated that the main objection to rate-distortion optimized coding, namely that it is too computationally demanding, can be mitigated by rate-distortion estimation. The strength of rate-distortion optimized coding lies in its flexibility, or specifically its adaptivity to user and/or channel constraints and the input signal; this will become an increasingly important factor in the future as the many different networks converge to one heterogeneous network. As we have already argued, the source coding problem will persist, but the relative weight between the different design criteria will change. We see already today that the bit-rate is not as important as it once was in mobile telephony.

References

[1] E. F. Schröder, "Digital audio broadcasting (DAB)," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 164–170.

[2] J. Herre and B. Grill, "Overview of MPEG-4 Audio and its applications in mobile communications," in *IEEE Int. Conf. Signal Processing*, vol. 1, 2000, pp. 11–20.

[3] F. Rumsey, "Putting low-bit-rate audio to work," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 155–163.

[4] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81(10), Oct. 1993.

[5] J. Cohen, "ISDN applications for bit-rate-reduced audio coding," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 171–181.

[6] G. C. P. Lokhoff, "The digital compact cassette," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 182–189.

[7] G. C. P. Lokhoff, "MiniDisc: Disc-based digital recodering for portable audio applications," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 190–196.

[8] C. E. Shannon, "A Mathematical Theory of Communication I," *The Bell Systems Technical Journal*, vol. 27, pp. 369–423, July 1948.

[9] C. E. Shannon, "A Mathematical Theory of Communication II," *The Bell Systems Technical Journal*, vol. 27, pp. 623–656, Oct. 1948.

[10] J. Lindblom, "Coding speech for packet networks," Ph.D. dissertation, Chalmers University of Technology, 2003.

[11] C. A. Rødbro, "Speech processing methods for the packet loss problem," Ph.D. dissertation, Aalborg University, 2004.

[12] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech and Audio Processing*, vol. 13(5), pp. 16–19, 2005.

[13] J. Lindblom and P. Hedelin, "Packet Loss Concealment Based on Sinusoidal Modeling," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, 2002, pp. 65–67.

[14] C. A. Rødbro, J. Jensen, and R. Heusdens, "Rate-distortion optimal time-segmentation and redundancy selection for VoIP," *IEEE Trans. Speech and Audio Processing*, 2005, accepted.

[15] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, "Hidden markov model based framework for packet loss concealment in voice over IP," *IEEE Trans. Speech and Audio Processing*, 2005, accepted.

[16] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Information Theory*, vol. 44(6), pp. 2693–2723, 1998.

[17] T. Berger, *Rate-distortion theory: A Mathematical basis for data compression*. Prentice-Hall, 1971.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[19] P. Kroon, "Evalation of speech coders," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 13.

[20] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88(4), pp. 451–515, Apr. 2000.

[21] F. Riera-Palou, A. C. den Brinker, and A. J. Gerrits, "A hybrid parametric-waveform approach to bistream scalable audio coding," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2250–2254.

[22] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1993.

[23] D. K. Neuhoff and R. M. Gray, "Quantization," in *IEEE Trans. Information Theory*, vol. 44, 1998.

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

[25] T. M. Apostol, *Mathematical Analysis*, 2nd ed. Addison-Wesley, 1974.

[26] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1445–1453, Sept. 1988.

[27] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech and Audio Processing*, pp. 646–655, 8(6) 2000.

[28] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE National Convention Records*, 1959, pp. 142–163.

[29] M. M. Goodwin, "Adaptive Signal Models: Theory, Algorithms, and Audio Applications," Ph.D. dissertation, University of California, Berkeley, 1997.

[30] M. M. Goodwin, "Multiscale overlap-add sinusoidal modeling using matching pursuit and refinements," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2001, pp. 207–210.

[31] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. Speech and Audio Processing*, vol. 1(4), pp. 386–399, Oct. 1993.

[32] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. Aud. Eng. Soc. 17th Conf.*, 1999, pp. 244–250.

[33] R. J. McAulay and T. F. Quatieri, "Speech Transformation Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1449–1464, Dec. 1986.

[34] R. Vafin and W. B. Kleijn, "Towards optimal quantization in multistage audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2004, pp. 205–208.

[35] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[36] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Prentice Hall, 1997.

[37] *JTC1/SC29/WG11 MPEG IS11172-3*, ISO/IEC Information technolgy - Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s–Part 3: Audio, 1992.

[38] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27(5), pp. 512–530, Oct. 1979.

[39] G. Schuller and J. Herre, "Audio coding: Recent advances and standards," Lecture Notes, Tutorial at IEEE Int. Conf. Acoust., Speech, and Signal Processing, 2004.

[40] M. Hans and R. W. Schafer, "Lossless compression of digital audio," *IEEE SP Mag.*, vol. 18(4), pp. 21–32, July 2001.

[41] P. Noll, "MPEG digital audio coding," *IEEE SP Mag.*, vol. 14(5), pp. 59–81, Sept. 1997.

[42] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2(2), pp. 60–74, 1995.

[43] A. S. Spanias, "Speech Coding: A Tutorial Review," in *Proc. IEEE*, vol. 82(10), Oct. 1994, pp. 1541–1582.

[44] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, vol. 82(6), pp. 900–918, June 1994.

[45] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: a theoretical survey," *J. Audio Eng. Soc.*, vol. 40, pp. 355–375, 1992.

[46] S. P. Lipshitz, J. Vanderkooy, and R. A. Wannamaker, "Minimally audible noise shaping," *J. Audio Eng. Soc.*, vol. 39, pp. 836–852, 1991.

[47] S. P. Lipshitz, R. A. Wannamaker, J. Vanderkooy, and J. N. Wright, "Non-subtractive dither," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 1991.

[48] R. A. Wannamaker, "Psycho-acoustically optimal noise-shaping," *J. Audio Eng. Soc.*, vol. 40, pp. 611–620, July/August 1992.

[49] R. A. Wannamaker, "The theory of dithered quantization," Ph.D. dissertation, University of Waterloo, 2003.

[50] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Prentice-Hall, 1984.

[51] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 881–884.

[52] G. D. T. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual audio coding using adaptive pre- and post-filters and lossless compression," in *IEEE Trans. Speech and Audio Processing*, vol. 10(6), Sept. 2002, pp. 379–390.

[53] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *Proc. 101st Conv. Aud. Eng. Soc.*, 1996, paper preprint 4384.

[54] M. Link, "An attack processing of audio signals for optimizing the temporal characteristics of a low bit-rate audio coding system," in *Proc. 95th Conv. Aud. Eng. Soc.*, 1993, paper preprint 3696.

[55] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, 2nd ed. Springer, 1999.

[56] K. Brandenburg, "Introduction to perceptual coding," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 23–30.

[57] B. C. J. Moore, "Masking in the human auditory system," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 9–19.

[58] R. Veldhuis and A. Kohlrausch, "Waveform coding and auditory masking," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 11, pp. 397–432.

[59] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1988, pp. 2524–2527.

[60] J. D. Johnston, "Transform coding of audio signal susing perceptual noise criteria," *IEEE J. Select. Areas Commun.*, pp. 314–323, 1988.

[61] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. Academic Press, 1997.

[62] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[63] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2001, pp. 1805 – 1808.

[64] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2004.

[65] 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. Am.*, vol. 99(6), pp. 3615–3622, June 1996.

[66] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. ii. simulations and measurements," *J. Acoust. Soc. Am.*, vol. 99(6), pp. 3623–3631, June 1996.

[67] J. Plasberg, D. Zhao, and W. B. Kleijn, "Sensitivity matrix for a spectro-temporal auditory model," in *Proc. European Signal Processing Conf.*, 2004, pp. 1673–1676.

[68] S. Voran, "Perception-based bit-allocation algorithms for audio coding," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 1997.

[69] F. Baumgarte, "Improved audio coding using a psychoacoustic model based on cochlear filter bank," *IEEE Trans. Speech and Audio Processing*, vol. 10(6), pp. 495–503, Oct. 2002.

[70] C. Colomes, M. Lever, J.-B. Rault, and Y. F. Dehery, "A perceptual model applied to audio bit-rate reduction," *J. Audio Eng. Soc.*, vol. 43, pp. 223–239, 1995.

[71] J. Breebaart, "Modeling binaural signal detection," Ph.D. dissertation, Technical University of Eindhoven, 2001.

[72] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24(5), pp. 380–391, 1976.

[73] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Trans. Speech and Audio Processing*, vol. 3(5), pp. 367–381, 1995.

[74] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, pp. 3–14, 1993.

[75] D. Sinha and A. H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Processing*, vol. 41, pp. 3463–3479, 1993.

[76] M. Erne, G. Moschytz, and C. Faller, "Best wavelet-packet bases for audio coding using perceptual and rate-distortion criteria," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1999, pp. 909–912.

[77] L. D. Fielder, M. Bosi, G. Davidson, M. Davis, C. todd, and S. Vernon, "AC-2 and AC3: Low-Complexity Transform-based Audio Coding," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 54–72.

[78] J. D. Johnston, D. Sinha, S. Dorward, and S. R. Quackenbush, "AT&T Perceptual Audio Coding (PAC)," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 73–82.

[79] F. Wylie, "apt-X100: Low-Delay, Low-Bit-Rate Subband ADPCM Digital Audio Coding," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 83–94.

[80] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, and R. M. Heddle, "ATRAC: Adaptive Transform Acoustic Coding for MiniDisc," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 95–101.

[81] K. Brandenburg and G. Stoll, "ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 31–42.

[82] *JTC1/SC29/WG11 MPEG IS13818-3*, ISO/IEC Information technolgy - Coding of moving pictures and associated audio–Part 3: Audio, 1994.

[83] *ISO/IEC IS13818-7*, ISO/IEC Information technology - generic coding of moving pictures and associated audio information. Part 7: Advanced Audio Coding, 1997.

[84] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc.*, vol. 45(10), pp. 789–814, 1997.

[85] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[86] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(5), pp. 1153–1161, Oct. 1986.

[87] J. Princen, A. Johnson, and A. Bradley, "Subband/transform coding using filter banks designs based on time domain aliasing cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1987, pp. 2161–2164.

[88] R. J. Beaton, J. G. Beerends, M. Keyhl, and W. C. Treurniet, "Objective perceptual measurement of audio quality," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 126–152.

[89] *ITU-T P.862*, ITU Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, Feb. 2001.

[90] *ITU-R BS.1387*, ITU Method for objective measurements of perceived audio quality, Nov. 2001.

[91] S. Voran, "Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique," *IEEE Trans. Speech and Audio Processing*, vol. 7(4), pp. 371–382, 1999.

[92] S. Voran, "Objective estimation of perceived speech quality. II. Evaluation of the measuring normalizing block technique," *IEEE Trans. Speech and Audio Processing*, vol. 7(4), pp. 383–390, 1999.

[93] *ITU-R BS.1534*, ITU Method for subjective assessment of intermediate quality level of coding system, 2001.

[94] *ITU-R P.800*, ITU Methods for Subjective Determination of Transmission Quality, Jan. 1996.

[95] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.

[96] T. Rydén, "Using listening tests to assess audio codecs," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 115–125.

[97] ISO/IEC JTC1/SC29/WG11, "Report on the formal subjective listening tests of MPEG-2 NBC multichannel audio coding, Tech. Rep. N1419, Nov. 1996.

[98] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4, pp. 121–174.

[99] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(4), pp. 744–754, Aug. 1986.

[100] P. Hedelin, "A tone oriented voice excited vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1981, pp. 205–208.

[101] E. B. George and M. J. T. Smith, "A new speech coding model based on a least-squares sinusoidal representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1987, pp. 1641–1644.

[102] J. S. Marques, I. M. Trancoso, and A. J. Abrantes, "Harmonic Coding of Speech: An Experimental Study," in *EuroSpeech Proceedings*, 1991, pp. 235–238.

[103] J. S. Marques and L. B. Almeida, "A Background for Sinusoid Based Representation of Voiced Speech," in *IEEE Trans. Acoust., Speech, Signal Processing*, 1986, pp. 1233–1236.

[104] J. S. Rodrigues and L. B. Almeida, "Harmonic Coding at 8kbits/sec," in *Proceedings IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1987, pp. 1621–1624.

[105] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech and Audio Processing*, vol. 5(5), pp. 389–406, Sept. 1997.

[106] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40(6), pp. 497–516, 1992.

[107] J. O. Smith III and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," in *Proceedings of the International Computer Music Conference (ICMC-87, Tokyo), Computer Music Association*, 1987.

[108] J. O. Smith III and X. Serra, "Spectral Modelling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, 1990.

[109] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates," in *Proc. 100th Conv. Aud. Eng. Soc.*, 1996, paper preprint 4179.

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

[111] H. Purnhagen and N. Meine, "HILN - The MPEG-4 Parametric Audio Coding Tools," in *IEEE International Symposium on Circuits and Systems*, 2000.

[112] B. Edler and H. Purnhagen, "Parametric audio coding," in *Proc. ICSP*, 2000, pp. 21–24.

[113] S. N. Levine, T. S. Verma, and J. O. Smith III, "Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 1997, pp. 101–104.

[114] S. N. Levine and J. O. Smith III, "A switched parametric & transform audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 985–988.

[115] K. N. Hamdy, M. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 1045–1048.

[116] T. S. Verma and T. H. Y. Meng, "A 6kbps to 85kbps scalable audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 877–880.

[117] A. C. den Brinker, E. G. P. Schuijers, and A. W. J. Oomen, "Parametric coding for high-quality audio," in *Proc. 112th Conv. Aud. Eng. Soc.*, 2002, paper preprint 5554.

[118] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and A. J. Gerrits, "Advances in parametric coding for high-quality audio," in *Proc 1st. IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA-2002)*, 2002.

[119] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *Proc. 114th Conv. Aud. Eng. Soc.*, 2003, paper preprint 5852.

[120] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 731–740, Oct. 2001.

[121] "Skype," June 2005, http://www.skype.com.

[122] T. S. Verma, "A perceptually based audio signal model with application to scalable audio compression," Ph.D. dissertation, Stanford University, 1999.

[123] J. Jensen, "Sinusoidal Models for Speech Signal Processing," Ph.D. dissertation, CPK, Institute of Electronic Systems, Aalborg University, 2000.

[124] L. Almeida and J. Tribolet, "Harmonic coding: A low bit-rate, good-quality speech coding technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 7, 1982, pp. 1664–1667.

[125] J. S. Marques, L. B. Almeida, and J. M. Tribolet, "Harmonic coding at 4.8kb/s," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1990, pp. 17–20.

[126] E. B. George and M. J. T. Smith, "Perceptual Conciderations in a Low Bit Rate Sinusoidal Vocoder," in *Ninth Annual International Phoenix Conference on Computers and Communications*, 1990, pp. 268–275.

[127] E. B. George and M. J. T. Smith, "Generalized overlap-add sinusoidal modeling applied to quasi-harmonic tone synthesis," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 1993, pp. 165–168.

[128] E. B. George and M. J. T. Smith, "Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model," in *IEEE Trans. Speech and Audio Processing*, vol. 3, 1997, pp. 389–406.

[129] J. Lattard, "Influence of inharmonicity on the tuning of a piano - measurements and mathematical simulation," *J. Acoust. Soc. Am.*, vol. 94, pp. 46–53, 1993.

[130] F. Myburg, "Design of a scalable parametric audio coder," Ph.D. dissertation, Technical University of Eindhoven, 2004.

[131] R. J. McAulay and T. F. Quatieri, "Multirate sinusoidal transform coding at rates from 2.4 kbps to 8 kbps," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1986, pp. 1645–1648.

[132] R. J. McAulay and T. F. Quatieri, "Phase modelling and its application to sinusoidal transform coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1986, pp. 1713–1715.

[133] L. B. Almeida and J. M. Tribolet, "A model for short-time phase prediction of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1981, pp. 213–216.

[134] S. Ahmadi and A. S. Spanias, "Improved algorithms for phase prediction and frame interpolation in low bit rate sinusoidal coders," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 1, 1998, pp. 362–366.

[135] S. Ahmadi and A. S. Spanias, "A new phase model for sinsoidal transform coding of speech," in *IEEE Trans. Speech and Audio Processing*, vol. 6(5), 1998, pp. 495–501.

[136] S. Ahmadi and A. S. Spanias, "New techniques for sinusoidal coding of speech at 2400 bps," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 1, 1996, pp. 495–501.

[137] A. C. den Brinker, A. J. Gerrits, and R. J. Sluijter, "Phase transmission in a sinusoidal audio and speech coder," in *Proc. 115th Conv. Aud. Eng. Soc.*, 2003, paper preprint 5983.

[138] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Pearson Professional Education, 2001.

[139] W. R. Gardner and B. D. Rao, "Mixed-phase ar models for voiced speech and perceptual cost functions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994, pp. 205–208.

[140] P. Hedelin, "Phase compensation in all-pole speech analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1988, pp. 339–342.

[141] S. Ahmadi, "An improved residual-domain phase/amplitude model for sinusoidal coding of speech at very low bit rates: A variable rate scheme," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 1999, pp. 2291–2294.

[142] J. Nieuwenhuijse, R. Heusdens, and E. F. Deprettere, "Robust Exponential Modeling of Audio Signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 3581–3584.

[143] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech and Audio Processing*, vol. 12(2), pp. 121–132, Mar. 2004.

[144] M. M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 2037–2040.

[145] R. Boyer and K. Abed-Meraim, "Audio Transient Modeling by Damped and Delayed Sinusoids (DDS)," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2002, pp. 1729–1732.

[146] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids," *IEEE Trans. Speech and Audio Processing*, vol. 12(2), pp. 110 – 120, Mar. 2004.

[147] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. van Huffel, "Perceptual audio modeling with exponentially damped sinusoids," *Signal Processing*, vol. 85, pp. 163–176, 2005.

[148] R. Heusdens, J. Jensen, W. B. Kleijn, V. kot, O. A. Niamut, S. van de Par, N. H. van Schijndel, and R. Vafin, "Bit-rate scalable intra-frame sinusoidal audio coding based on rate-distortion optimisation," *J. Audio Eng. Soc.*, 2005, submitted.

[149] T. D. Rossing, *The Science of Sound*, 2nd ed. Addison-Wesley Publishing Company, 1990.

[150] *ISO/IEC 14496-3:2001/AMD2*, ISO/IEC Parametric Coding for High-Quality Audio, July 2004.

[151] F. Myburg, A. C. den Brinker, and S. van Eijndhoven, "Sinusoidal analysis of audio with polynomial phase and amplitude," in *Proc. ProRISC*, 2001.

[152] F. Myburg, A. C. den Brinker, and S. van Eijndhoven, "Multi-component chirp analysis in parametric audio coding," in *Fourth IEEE Benelux Signal Processing Symposium*, 2004.

[153] G. Li, L. Qiu, and L. K. Ng, "Signal representation based on instantaneous amplitude models with application to speech synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 8(3), pp. 353–357, 2000.

[154] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Trans. Signal Processing*, vol. 48(2), pp. 338–352, Feb. 2000.

[155] A.-J. van der Veen, E. F. Deprettere, and A. L. Swindlehurst, "Subspace-based signal analysis uisng singular value decompostion," *Proc. IEEE*, vol. 81(9), pp. 1277–1308, Sept. 1993.

[156] H. Krim and M. Viberg, "Two decades of array signal processing research–the parametric approach," *IEEE SP Mag.*, July 1996.

[157] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 306–309.

[158] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34(3), pp. 276–280, Mar. 1986.

[159] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophys. J. Roy. Astron. Soc.*, vol. 33, pp. 347–366, 1973.

[160] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(7), July 1989.

[161] P. Stoica, A. Jakobsson, and J. Li, "Cisiod parameter estimation in the coloured noise case: Asymptotic cramér-rao bound, maximum likelihood, and nonlinear least-squares," in *IEEE Trans. Signal Processing*, vol. 45(8), Aug. 1997, pp. 2048–2059.

[162] P. Stoica, R. L. Moses, B. Friedlander, and T. Söderström, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(3), pp. 378–392, Mar. 1989.

[163] M. M. Goodwin, "Matching pursuit and atomic signal models based on recursive filter banks," *IEEE Trans. Signal Processing*, vol. 47(7), pp. 1890–1902, July 1999.

[164] R. Vafin, S. V. Andersen, and W. B. Kleijn, "Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2000, pp. 901–904.

[165] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[166] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1999, pp. 981–984.

[167] H. Purnhagen, N. Meine, and B. Edler, "Sinusoidal coding using loudness-based component selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2002, pp. 1817–1820.

[168] T. Painter and A. S. Spanias, "Perceptual segmentation and component selection in compact sinusoidal representation of audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, 2001, pp. 3289–3292.

[169] T. Painter and A. Spanias, "Perceptual segmentation and component selection for sinusoidal representations of audio," *IEEE Trans. Speech and Audio Processing*, vol. 13(2), pp. 149–162, Mar. 2005.

[170] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1982.

[171] R. Vafin and W. B. Kleijn, "Entropy-constrained polar quantization: Theory and an application to audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2002, pp. 1837–1840.

[172] R. Vafin, D. Prakash, and W. B. Kleijn, "On frequency quantization in sinusoidal audio coding," *IEEE Signal Processing Lett.*, vol. 12, no. 3, pp. 210–213, Mar. 2005.

[173] R. Vafin and W. B. Kleijn, "Entropy-constrained polar quantization and its application to audio coding," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 220–232, Mar. 2005.

[174] R. Vafin, "Towards flexible audio coding," Ph.D. dissertation, Royal Institute of Technology, Dec. 2004, tRITA-S3-SIP-2004-4.

[175] R. Heusdens and J. Jesper, "Jointly optimal time segmentation, component selection and quantization for sinusoidal coding of audio and speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2005, pp. 18–23.

[176] P. E. L. Korten, J. Jensen, and R. Heusdens, "High resolution spherical quantization of sinusoidal parameters uinsg a prceptual distortion measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2005, pp. 177–180.

[177] P. E. L. Korten, J. Jensen, and R. Heusdens, "High rate spherical quantization of sinusoidal parameters," in *Proc. European Signal Processing Conf.*, 2004, pp. 1805–1808.

[178] J. Lindblom and P. Hedelin, "Variable-dimension quantization of sinusoidal amplitudes using gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2004, pp. 100–103.

[179] J. Jensen and R. Heusdens, "A comparison of differential schemes for low-rate sinusoidal audio coding," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2003, pp. 205–208.

[180] J. Jensen and R. Heusdens, "Optimal frequency-differential encoding of sinusoidal model parameters," in *Proc. IEEE Conf. on Acoust., Speech, Signal Processing*, 2002, pp. 2497–2500.

[181] J. Jensen and R. Heusdens, "Schemes for optimal frequency-differential encoding of sinusoidal model parameters," *Elsevier Science Signal Processing*, vol. 83(8), pp. 1721–1735, Aug. 2003.

[182] R. Vafin and W. B. Kleijn, "Rate-distortion optimized quantization in multistage audio coding," *IEEE Trans. Speech and Audio Processing*, 2004, accepted.

[183] P. Kroon, E. D. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective multipulse coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1054–1063, 1986.

[184] R. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2004, pp. 189–192.

[185] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, 1989.

[186] M. M. Goodwin, "Nonuniform filterbank design for audio signal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1997, pp. 1229–1233.

[187] M. M. Goodwin, "Residual modeling in music analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1996, pp. 1005–1008.

[188] T. S. Verma and T. H. Y. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, 1998, pp. 3573–3576.

[189] S. N. Levine, "Audio Representations For Data Compression And Compressed Domain Processing," Ph.D. dissertation, Stanford University, 1999.

[190] D. Schulz, "Improving audio codecs by noise substitution," *J. Audio Eng. Soc.*, vol. 7/8, pp. 593–598, Jul/Aug 1996.

[191] J. Herre and D. Schulz, "Extending the MPEG-4 AAC codec by Perceptual Noise Substitution," in *Proc. 104th Conv. Aud. Eng. Soc.*, 1998, paper preprint 4720.

[192] A. Härmä, "Frequency-warped autoregressive modeling and filtering," Ph.D. dissertation, Helsinki University of Technology, 2001.

[193] A. Härmä, U. K. Laine, and M. Karjalainen, "Warped linear prediction (WLP) in audio coding," in *Nordic Signal Processing Symposium*, 1996.

[194] A. C. den Brinker, V. Voitischchuk, and S. J. L. van Eijndhoven, "IIR-based pure linear prediction," *IEEE Trans. Speech and Audio Processing*, vol. 12(1), Jan. 2004.

[195] F. K. Soong and B. Juang, "Optimal quantization of lsp parameters," in *IEEE Trans. Speech and Audio Processing*, vol. 1, 1993, pp. 15–24.

[196] W. B. Kleijn and K. K. Paliwal, "Quantization of LPC Parameters," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 12.

[197] O. A. Niamut and R. Heusdens, "RD Optimal Time Segmentation for the Time-Varying MDCT," in *Proc. European Signal Processing Conf.*, Sept. 2004, pp. 1649–1652.

[198] O. A. Niamut and R. Heusdens, "Flexible frequency decompositions for cosine-modulated filter banks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 2003, pp. 449–452.

[199] P. Prandoni, "Optimal Segmentation Techniques for Piecewise Stationary Signals," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne, 1999.

[200] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[201] B. Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," *Frequenz*, pp. 1033–1036, 1989.

[202] K. Brandenburg, J. Herre, J. D. Johnston, Y. Mahieux, and E. F. Schroeder, "ASPEC: Adaptive Spectral Entropy Coding of High Quality Music Signals," in *90th Conv. Aud. Eng. Soc.*, 1991, paper Preprint 3011 A-4.

[203] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1992, pp. 569–572.

[204] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[205] P. Prandoni, M. M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 2029–2032.

[206] K. Engan, "Frame based signal representation and compression," Ph.D. dissertation, Stavanger Unversity College, 2000.

[207] K. Engan, S. O. Aase, and J. H. Husøy, "Designing frames for matching pursuit algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 1817–1820.

[208] M. Dietz, L. Liljeryd, K. Kjörling, and U. Kunz, "Spectral band replication, a novel approach to audio coding," in *Proc. 112th Conv. Aud. Eng. Soc.*, 2002, paper preprint 5553.

[209] J. Herre, "From joint stereo to spatial audio coding–recent progress and standardization," in *Proc. Int. Conf. Digital Audio Effects*, 2004, pp. 157–162.

[210] F. Baumgarte and C. Faller, "Binaural Cue Coding–Part I: Psychocaoustic fundamentals and design principles," *IEEE Trans. Speech and Audio Processing*, vol. 11(6), pp. 509–519, 2003.

[211] C. Faller and F. Baumgarte, "Binaural Cue Coding–Part II: Schemes and applications," *IEEE Trans. Speech and Audio Processing*, vol. 11(6), pp. 520–531, 2003.

[212] R. van de Waal and R. Velduis, "Subband coding of stereophonic digital audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 3601–3604.

[213] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *Proc. 96th Conv. Aud. Eng. Soc.*, 1994, paper Preprint 3799.

[214] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "High-quality parametric spatial audio coding at low bit rates," in *Proc. 116th Conv. Aud. Eng. Soc.*, 2004, paper Prerpint 6072.

[215] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1305–1322, June 2005.

[216] G. Stoll, "ISO-MPEG-2 Audio: A Generic Standard for the Coding of Two-Channel and Multichannel Sound," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 43–53.

[217] *Federal Standard 1015*, National Communication System–Office of Technology and Standards Telecommunications: Analog to digital conversion of radio voice by 2400 bit/second linear predictive coding, national communication system, Nov. 1984.

[218] E. Bedrosian, "A product theorem for Hilbert transforms," in *Proc. IEEE*, vol. 51(1), May 1963, pp. 868–869.

[219] S. L. Hahn, *Hilbert Transforms in Signal Processing*. Artech House, 1996.

[220] F. Nordén, S. V. Andersen, S. H. Jensen, and W. B. Kleijn, "Property vector based distortion estimation," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2275–2279.

Paper A

Amplitude Modulated Sinusoidal Models for Audio Modeling and Coding

Mads Græsbøll Christensen, Søren Vang Andersen, and
Søren Holdt Jensen

The paper has been published in
Knowledge-Based Intelligent Information & Enginerring Systems, V. Palade,
R. J. Howlett, and L. C. Jain, Eds., *Lecture Notes in Artificial Intelligence*,
Springer-Verlag, vol. 2773, pp. 1334–1342, 2003.

Note

The signal model $\hat{x}(n)$ and the subband signal $x_q(n)$ only relate to the amplitude modulating signal as described in (6) and (7) for the real case, i.e. $\gamma_q^*(n) = \gamma_q(n)$. For the complex case, (6) should read

$$\hat{x}(n) = \sum_{q=1}^{Q} \gamma_q^*(n) \frac{A_q}{2} \exp(-j(\omega_q + \phi_q)) + \gamma_q(n) \frac{A_q}{2} \exp(j(\omega_q + \phi_q)).$$

Similarly, the subband signal in (7) can be written as

$$\begin{aligned} x_q(n) &= \gamma_q^*(n) \frac{A_q}{2} \exp(-j(\omega_q + \phi_q)) + \gamma_q(n) \frac{A_q}{2} \exp(j(\omega_q + \phi_q)) \\ &= \gamma_q^*(n) s_q^*(n) + \gamma_q(n) s_q(n), \end{aligned}$$

where $s_q(n) = \exp(j(\omega_q + \phi_q))$. With these definitions in place, the remaining part of the section holds for both the complex and the real case.

The layout has been revised.

Abstract

In this paper a new perspective on modeling of transient phenomena in the context of sinusoidal audio modeling and coding is presented. In our approach the task of finding time-varying amplitudes for sinusoidal models is viewed as an AM demodulation problem. A general perfect reconstruction framework for amplitude modulated sinusoids is introduced and model reductions lead to a model for audio compression. Demodulation methods are considered for estimation of the time-varying amplitudes, and inherent constraints and limitations are discussed. Finally, some applications are considered and discussed and the concepts are demonstrated to improve sinusoidal modeling of audio and speech.

1 Introduction

In the last couple of decades sinusoidal modeling and coding of both speech and audio in general has received great attention in research. In its most general form, it models a segment of a signal as a finite sum of sinusoidal components each having a time-varying amplitude and a time-varying instantaneous phase. Perhaps the most commonly used derivative of this model is the constant-frequency constant-amplitude model known as the basic sinusoidal model. This model is based on the assumptions that the amplitudes and frequencies remain constant within the segment. It has been used for many years in speech modeling and transformation [1]. The model, however, has problems in modeling transient phenomena such as onsets, which causes so-called pre-echos to occur. This is basically due to the quasi-stationarity assumptions of the model being violated and the fundamental trade-off between time and frequency resolution. Also, the use of overlap-add or interpolative synthesis inevitably smears the time-resolution.

Many different strategies for handling time-varying amplitudes have surfaced in recent years. For example, the use of time-adaptive segmentation [2] improves performance greatly at the cost of increased delay. But even then pre-echos may still occur in overlap regions or if interpolative synthesis [1] is used. Also, the use of exponential dampening of each sinusoid has been extensively studied [3–5], although issues concerning quantization remain unsolved. Other approaches include the use of one common dampening factor for all sinusoids [6], the use of asymmetric windows [7], the use of an envelope estimated by low-pass filtering of the absolute value of the input [8] and the approaches taken in [9, 10]. In [9] lines are fitted to the instantaneous envelope and then used in sinusoidal modeling, and in [10] transient locations are modified in time to reduce pre-echo artifacts. The latter requires the use of dynamic time segmentation. Also, tracking of individual speech formants by means of an energy separation into amplitude modulation and frequency modulation (FM) contributions has been studied in [11–13].

In this paper we propose amplitude modulated sinusoidal models for audio model-

ing and coding applications. The rest of the paper is organized as follows: In section 2 the mathematical background is presented. A general perfect reconstruction model is derived in section 3, and in section 4 a model which addresses one of the major issues of audio coding regardless of type, namely pre-echo, is presented along with a computationally simple estimation technique. Finally, in section 5 some experimental results are presented and discussed and section 6 concludes on the work.

2 Some Preliminaries

The methods proposed in this paper are all based on the so-called analytic signal, which is derived from the Hilbert transform. First, we introduce the Hilbert transform and define the analytic signal and the instantaneous envelope. Then we briefly state Bedrosian's theorem, which is essential to this paper.

Definition 1 (Discrete Hilbert Transform). Let $x_r(n)$ be a discrete-time real signal. The Discrete Hilbert transform, $\mathcal{H}\{\cdot\}$, of this, denoted $x_i(n)$, is then defined as (see e.g. [14])

$$x_i(n) = \mathcal{H}\{x_r(n)\} = \sum_{m=-\infty}^{\infty} h(m)x_r(n-m). \tag{1}$$

where $h(n)$ is the impulse response of the discrete Hilbert transform given by

$$h(n) = \begin{cases} \frac{2\sin^2(\pi n/2)}{\pi n}, & n \neq 0 \\ 0, & n = 0 \end{cases} . \tag{2}$$

A useful way of looking at the Hilbert transform, and perhaps a more intuitive definition, is in the frequency domain:

$$X_i(\omega) = H(\omega)X_r(\omega), \quad \text{with} \quad H(\omega) = \begin{cases} j, & \text{for} \quad -\pi < \omega < 0 \\ 0, & \text{for} \quad \omega = \{0, \pi\} \\ -j, & \text{for} \quad 0 < \omega < \pi \end{cases} . \tag{3}$$

where $X_i(\omega)$ and $X_r(\omega)$ are the Fourier transform (denoted $\mathcal{F}\{\cdot\}$) of $x_i(n)$ and $x_r(n)$, respectively, and $H(\omega)$ is the Fourier transform of $h(n)$. The so-called analytic signal and instantaneous envelope are then defined as

$$x_c(n) = x_r(n) + jx_i(n) \quad \text{and} \quad |x_c(n)| = \sqrt{x_r^2(n) + x_i^2(n)}, \tag{4}$$

respectively. With these definitions in place, we now state Bedrosian's theorem [15].

Theorem 1 (Bedrosian). *Let $f(n)$ and $g(n)$ denote generally complex functions in $\ell^2(\mathbb{Z})$ of the real, discrete variable n. If*

1. *the Fourier transform $F(\omega)$ of $f(n)$ is zero for $a < |\omega| \leq \pi$ and the Fourier transform $G(\omega)$ of $g(n)$ is zero for $0 \leq |\omega| < a$, where a is an arbitrary positive constant, or*

2. *$f(n)$ and $g(n)$ are analytic, then*

$$\mathcal{H}\{f(n)g(n)\} = f(n)\mathcal{H}\{g(n)\}. \tag{5}$$

For proof of the continuous case see [15]. The theorem holds also for periodic signals in which case the Fourier series should be applied.

3 Sum of Amplitude Modulated Sinusoids

In this section we consider a perfect reconstruction framework based on a model consisting of a sum of amplitude modulated sinusoids:

$$\hat{x}(n) = \sum_{q=1}^{Q} \gamma_q(n) A_q \cos(\omega_q n + \phi_q) \quad \text{for} \quad n = 0, \ldots, N-1, \tag{6}$$

where $\gamma_q(n)$ is the amplitude modulating signal, A_q the amplitude, ω_q the frequency, and ϕ_q the phase of the qth sinusoid. We note in the passing that the aforementioned exponential sinusoidal models [3–5] fall into this category. Assume that the signal has been split into a set of subbands by a perfect reconstruction nonuniform Q-band filterbank, such as [16], having a set of cut-off frequencies Ω_q for $q = 0, 1, \ldots, Q$ where $\Omega_0 = 0$ and $\Omega_Q = \pi$. Then we express the contents of each individual subband $x_q(n)$ as an amplitude modulated sinusoid placed in the middle of the band, i.e.

$$x_q(n) = \gamma_q(n) A_q \cos(\omega_q n + \phi_q) = \gamma_q(n) s_q(n), \tag{7}$$

where $\omega_q = \frac{\Omega_q + \Omega_{q-1}}{2}$, $\gamma_q(n) \in \mathbb{C}$, i.e. the modulation is complex. We start our demodulation by finding the analytic signal representation of both the left and right side of the previous equation:

$$x_q(n) + j\mathcal{H}\{x_q(n)\} = \gamma_q(n) s_q(n) + j\mathcal{H}\{\gamma_q(n) s_q(n)\}, \tag{8}$$

which according to Bedrosian's theorem is equal to

$$\gamma_q(n) s_q(n) + j\mathcal{H}\{\gamma_q(n) s_q(n)\} = \gamma_q(n) \left(s_q(n) + j\mathcal{H}\{s_q(n)\}\right) \tag{9}$$

$$= \gamma_q(n) A_l \exp\left(j(\omega_q n + \phi_q)\right). \tag{10}$$

This means that we can simply perform complex demodulation in each individual subband using a complex sinusoid, i.e.

$$\gamma_q(n) = \left(x_q(n) + j\mathcal{H}\{x_q(n)\}\right) \frac{1}{A_q} \exp(-j(\omega_q n + \phi_q)). \tag{11}$$

In this case we have a modulation with a bandwidth equal to the bandwidth of the subband, $\Delta_q = \Omega_q - \Omega_{q-1}$. It is of interest to relax the constraint on the frequency of the carrier. Here we consider a more general scenario, where the carrier may be placed anywhere in the subband, i.e. $\Omega_{q-1} \leq \omega_q \leq \Omega_q$. In this case, the modulation is asymmetrical around the carrier in the spectrum. An alternative interpretation is that the carrier is both amplitude and phase modulated simultaneously. Alternatively, we can split the modulation into an upper (usb) and a lower sideband (lsb). These can be obtained by calculating the analytic signal of $\gamma_q(n)$ and $\gamma_q^*(n)$, which is similar to zeroing out the negative frequencies:

$$\gamma_{q,usb}(n) = \frac{1}{2}\left(\gamma_q(n) + j\mathcal{H}\{\gamma_q(n)\}\right) \tag{12}$$

$$\gamma_{q,lsb}(n) = \frac{1}{2}\left(\gamma_q^*(n) + j\mathcal{H}\{\gamma_q^*(n)\}\right). \tag{13}$$

The complex modulating signal can be reconstructed as

$$\gamma_q(n) = \gamma_{q,usb}(n) + \gamma_{q,lsb}^*(n). \tag{14}$$

The modulating signal can be written as $\gamma_q(n) = C + b(n)$, where $b(n)$ is zero mean. For $C \neq 0$, this is the case where the sinusoidal carrier is present in the spectrum in the form of a discrete frequency component. For the special case that $C = 0$, we have what is known as suppressed carrier AM, i.e. the carrier will not be present in the spectrum. In the context of speech modeling this representation may be useful in modeling non-tonal parts, e.g. unvoiced speech, whereas the non-suppressed AM ($C \neq 0$) case may be well-suited for voiced speech. In the particular case that the modulating signal is both non-negative and real, i.e. $\gamma_q(n) \in \mathbb{R}$ and $\gamma_q(n) \geq 0$, the demodulation simply reduces to

$$\gamma_q(n) = \frac{1}{A_q}|x_q(n) + j\mathcal{H}\{x_q(n)\}|, \tag{15}$$

as the instantaneous envelope of the carrier is equal to 1. This last estimation is lossy as opposed to the previous demodulations. Notice that in the perfect reconstruction scenario, the filtering of the signal into subbands and subsequent demodulation can be implemented efficiently using an FFT.

An alternative to the filterbank-based sum of amplitude modulated sinusoids scheme, which requires that the sinusoidal components are well spaced in frequency is the use of periodic algebraic separation [17, 18]. This allows for demodulation of closely spaced periodic components provided that the periods are known.

4 Amplitude Modulated Sum of Sinusoids

In this section a model for audio compression is introduced. This model addresses one of the major problems of audio coding regardless of type, namely pre-echo control. The

perfect reconstruction model of the previous section has an amplitude modulating signal of each individual sinusoid. Here, we explore the notion of having more sinusoids in each subband and that modulating signal being identical for all sinusoids in the subband. This is especially useful in the context of modeling onsets and may even be used in the one-band case for low bit-rate or single source applications. The model of the qth subband is:

$$\hat{x}_q(n) = \gamma_q(n) \sum_{l=1}^{L_q} A_{q,l} \cos(\omega_{q,l} n + \phi_{q,l}) = \gamma_q(n) \hat{s}_q(n), \tag{16}$$

where $\hat{s}_q(n)$ is the constant-amplitude part. In the one-band case where $x_q(n) = x(n)$ the models in [6–9] all fall into this category. These, however, do not reflect human sound perception very well as pre-echos may occur in the individual critical bands (see e.g. [19]). Neither do they take the presence of multiple temporally overlapping sources into account. The sum of amplitude modulated sinusoids, however, does take multiple sources into account.

The basic principle in the estimation of the modulating signal $\gamma_q(n)$ is that it can be separated from the constant-amplitude part of our model $\hat{x}_q(n)$ under certain conditions. First we write the instantaneous envelope of equation 16, i.e.

$$|\hat{x}_q(n) + j\mathcal{H}\{\hat{x}_q(n)\}| = |\gamma_q(n)\hat{s}_q(n) + j\mathcal{H}\{\gamma_q(n)\hat{s}_q(n)\}|. \tag{17}$$

Since we are concerned here with sinusoidal modeling, we constrain the modulation to the case of non-suppressed carrier and the physically meaningful non-negative and real modulating signal. Equation (17) can then be rewritten using Bedrosian's theorem:

$$|\gamma_q(n)\hat{s}_q(n) + j\mathcal{H}\{\gamma_q(n)\hat{s}_q(n)\}| = \gamma_q(n)|\hat{s}_q(n) + j\mathcal{H}\{\hat{s}_q(n)\}|. \tag{18}$$

For this to be true, our constant-amplitude model and the amplitude modulation must not overlap in frequency, i.e. we have that the lowest frequency must be above the bandwidth, BW, of the modulating signal

$$BW < \min_l \omega_{q,l}. \tag{19}$$

Using this constraint, we now proceed in the estimation of the amplitude modulating signal $\gamma_q(n)$ by finding the analytic signal of the sinusoidal model

$$\hat{x}_{q,c}(n) = \sum_{l=1}^{L_q} A_l \gamma_q(n) \exp(j\phi_{q,l}) \exp(j\omega_{q,l} n), \tag{20}$$

with subscript c denoting the analytic signal. We then find the squared instantaneous envelope of the model:

$$|\hat{x}_{q,c}(n)|^2 = \sum_{l=1}^{L_q} \sum_{k=1}^{L_q} \gamma_q^2(n) A_{q,l} A_{q,k} \exp(j(\phi_{q,k} - \phi_{q,l})) \exp(j(\omega_{q,k} - \omega_{q,l})n). \tag{21}$$

The squared instantaneous envelope is thus composed of a set of auto-terms ($l = k$) which identifies the amplitude modulating signal and a set of interfering cross-terms ($l \neq k$). From this it can be seen that the frequencies of these cross-terms in the instantaneous envelope is given by the distances between the sinusoidal components. Thus, the lowest frequency in the squared instantaneous envelope caused by the interaction of the constant-amplitude sinusoids is given by the minimum distance between two adjacent sinusoids. A computationally simple approach is to reduce the cross-terms by constraining the minimum distance between sinusoids and then simply lowpass filtering the squared instantaneous envelope of the input signal, i.e.

$$\gamma_q^2(n) = \alpha e_q^2(n) * h_{LP}(n), \tag{22}$$

where $e_q^2(n) = x_q^2(n) + \mathcal{H}\{x_q(n)\}^2$, α is a positive scaling factor and $h_{LP}(n)$ is the impulse response of an appropriate lowpass filter with a stopband frequency below half the minimum distance between two sinusoids, i.e.

$$2BW < \min_{l \neq k} |\omega_{q,l} - \omega_{q,k}|. \tag{23}$$

This estimate allows us to find an amplitude modulating signal without knowing the parameters of the sinusoidal model a priori. This is especially attractive in the context of matching pursuit [20]. Note that the constraint in equation (23) is more restrictive than those of theorem 1. The design of the lowpass filter is subject to conflicting criteria. On one hand, we want to have sufficient bandwidth for modeling transients well. On the other, we want to attenuate the cross-terms while having arbitrarily small spacing in frequency between adjacent sinusoids. Also these criteria have a time-varying nature. A suitable filter which can easily be altered to fit the requirements is described in [8]. Generally, the consequences of setting the cutoff frequency of the lowpass filter too low are more severe than setting it too high. Setting the cutoff frequency too high causes cross-terms to occur in $\gamma_q(n)$, which may result in degradation in some cases, whereas setting the cutoff frequency too low reduces the model's ability to handle transients.

An alternative approach in finding $\gamma_q(n)$ would be to estimate the amplitude modulating signals of the individual sinusoids and then combine these according to frequency bands or sources.

5 Results and Discussion

The framework in section 3 has been verified in simulations to attain perfect reconstruction. The choice of model, whether it is some derivative of the sum of amplitude modulated sinusoids or the amplitude modulated sum of sinusoids, should reflect signal characteristics. Types of sinusoidal signals that can be efficiently modeled using a one-band amplitude modulated sum of sinusoids are single sources that have a quasi-harmonic structure, i.e. pitched sounds. For example, voiced speech can be modeled well using

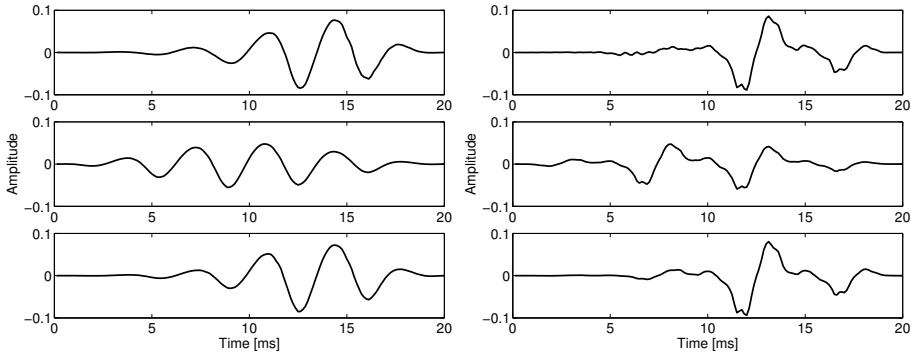

Figure 1: Signal examples: voiced speech. Original (top), modeled without AM (middle) and with AM (bottom).

such a model. In figure 1 two examples of onsets of voiced speech are shown (sampled at 8 kHz) with the originals at the top, modeled without AM in the middle, and with at the bottom. The fundamental frequency was found using a correlation-based algorithm and the amplitudes and phases were then estimated using weighted least-squares. Segments of size 20 ms and overlap-add with 50% overlap was used. It can be seen that the pre-echo artifacts present in the constant-amplitude model are clearly reduced by the use of the AM scheme. The proposed model and estimation technique was found to consistently improve performance of the harmonic sinusoidal model in transient speech segments with pre-echo artifacts clearly being reduced.

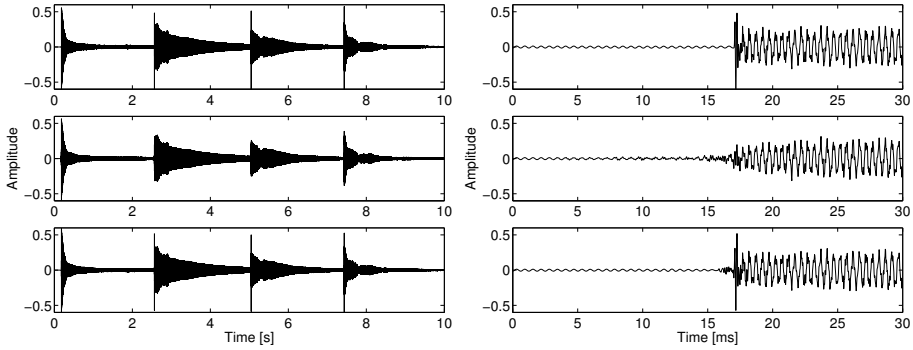

Figure 2: Signal examples: glockenspiel. Original (top), modeled without AM (middle) and with AM (bottom).

More complex signals composed of multiple temporally overlapping sources, however, require more sophisticated approaches for handling non-stationarities. The glock-

enspiel of SQAM [21] is such a signal. At first glance this signal seems well suited for modeling using a sinusoidal model. The onsets are, however, extremely difficult to model accurately using a sinusoidal model. This is illustrated in figure 2, again with the original at the top, modeled using constant amplitude sinusoids in the middle and using AM at the bottom. The signal on the left is the entire signal and the signal on the right is a magnification of a transition region between notes. In this case amplitude modulation is applied per equivalent rectangular bandwidth (ERB) (see [19]) and a simple matching pursuit-like algorithm was used for finding sinusoidal model parameters, i.e. no harmonic constraints on the frequencies. Again overlap-add using segments of 20 ms and 50% overlap was employed. In this example the sampling frequency was 44.1 kHz. It can be seen that the onsets are smeared when employing constant amplitude and that there is a significant improvement when AM is applied, although some smearing of the transition still occurs due to the filtering.

6 Conclusion

In this paper we have explored the notion of amplitude modulated sinusoidal models. First, a general perfect reconstruction framework based on a filterbank was introduced, and different options with respect to modulation and their physical interpretations were presented. Here, one sinusoid per subband is used and everything else in the subband is then modeled as modulation of that sinusoid. This model is generally applicable and can be used for modeling not only tonal signals but also noise-like signals such as unvoiced speech. Then a physically meaningful, compact representation for sinusoidal audio coding and modeling and a demodulation scheme with low computational complexity was presented. In this model, each subband is represented using a sum of sinusoids having one common real, non-negative modulating signal, which is estimated by lowpass filtering the squared instantaneous envelope. The model and the proposed estimation technique was found to be suitable for modeling of onsets of pitched sounds and was verified to generally improve modeling performance of sinusoidal models.

References

[1] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(4), pp. 744–754, Aug. 1986.

[2] P. Prandoni, M. M. Goodwin, and M. Vetterli, “Optimal time segmentation for signal modeling and compression,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 2029–2032.

[3] M. M. Goodwin, “Matching pursuit with damped sinusoids,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 2037–2040.

[4] J. Nieuwenhuijse, R. Heusdens, and E. F. Deprettere, "Robust Exponential Modeling of Audio Signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 3581–3584.

[5] J. Jensen, S. H. Jensen, and E. Hansen, "Exponential Sinusoidal Modeling of Transitional Speech Segments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 473–476.

[6] J. Jensen, S. H. Jensen, and E. Hansen, "Harmonic Exponential Modeling of Transitional Speech Segments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 1439–1442.

[7] R. Gribonval, P. Depalle, X. Rodet, E. Bacry, and S. Mallat, "Sound signal decomposition using a high resolution matching pursuit," in *Proc. Int. Computer Music Conf.*, Aug. 1996, pp. 293–296.

[8] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40(6), pp. 497–516, 1992.

[9] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates," in *100th Conv. Aud. Eng. Soc.*, 1996, paper preprint 4179.

[10] R. Vafin, R. Heusdens, and W. B. Kleijn, "Modifying transients for efficient coding of audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 3285–3288.

[11] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Processing*, vol. 41(10), pp. 3024–3051, Oct. 1993.

[12] A. C. Bovik, J. P. Havlicek, M. D. Desai, and D. S. Harding, "Limits on discrete modulated signals," *IEEE Trans. Signal Processing*, vol. 45(4), pp. 867–879, Apr. 1997.

[13] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary, "AM-FM Separation using Auditory-motivated Filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, 1996, pp. 465–480.

[14] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 1st ed. Prentice-Hall, 1989.

[15] E. Bedrosian, "A product theorem for Hilbert transforms," in *Proc. IEEE*, vol. 51(1), May 1963, pp. 868–869.

[16] M. M. Goodwin, "Nonuniform filterbank design for audio signal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1997, pp. 1229–1233.

[17] M.-Y. Zou, C. Zhenming, and R. Unbehauen, "Separation of periodic signals by using an algebraic method," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 5, 1991, pp. 2427–2430.

[18] B. Santhanam and P. Maragos, "Multicomponent AM-FM demodulation via periodicity-based algebraic separation and energy-based demodulation," *IEEE Trans. Communcations*, vol. 48(3), pp. 473–490, 2000.

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. Academic Press, 1997.

[20] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[21] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.

Paper B

Multiband Amplitude Modulated Sinusoidal Audio Modeling

Mads Græsbøll Christensen, Steven van de Par, Søren Holdt Jensen, and Søren Vang Andersen

The paper has been published in
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 169–172, 2004.

Abstract

In this paper, we investigate the importance of taking frequency-dependent temporal phenomena into account in audio coding. We do this in the context of sinusoidal modeling of audio signals by applying amplitude modulation to the sinusoidal components. Traditionally, audio coders use a fixed time-segmentation for all frequencies despite that it is well-known that the time-frequency resolution of the human auditory system is not constant. The well-known window switching is an example of this. We compare multiband amplitude modulated sinusoidal models to a singleband model using different audio excerpts. Based on both comparative listening tests and a psychoacoustical distortion measure it is concluded that an improvement is generally gained using multiband amplitude modulation, although specific single sources are well-modeled using a singleband model.

1 Introduction

A well-known problem in perceptual audio coding and modeling is what is known as pre-echo distortion or pre-echos (see e.g. [1]). Pre-echos can be defined as the introduction of a modeling error or quantization error that occurs before a transient signal. These occur in block-based modeling when there is an onset or attack at the end of a segment.

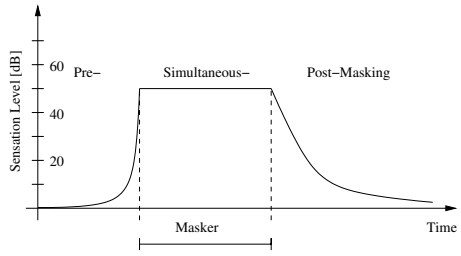

Figure 1: Temporal or nonsimultaneous masking properties. Pre-masking occurs before the onset of the masker and post-masking occurs afterwards (after [2]).

The importance of pre-echo control in audio coding and modeling can be understood by considering the temporal masking properties of the human auditory system. In audio coding the original signal serves as a masker of the error-signal. This masking is very effective when the error-signal is simultaneous with, or directly follows the masker. However, when the error-signal preceeds the masker, very little masking is observed. This is depicted in Figure 1 showing masking thresholds as a function of time. Pre-masking can be measured to typically last only about 20 ms, whereas post-masking can last longer than 100 ms [2]. Trained listeners, however, may exhibit little or no

pre-masking except for very short signals [3]. This means that any artifacts introduced before an onset are very poorly masked compared to the situation where a signal is present. The point that motivates this work is that masking phenomena occur on a critical band basis. Singleband techniques such as window switching [4] or AM as used in [5] do not take this into account. What may happen is that the choice of window length or the estimation of the amplitude modulating signal may be dominated by a stationary low-frequency component while a transient occurs at high frequencies, whereby audible artifacts are caused. Or it may happen that a short window is chosen because of some high-frequency transient while stationary low-frequency parts may suffer because of the decreased frequency resolution.

In sinusoidal coding of speech [6] and audio, fixed segmentation for all frequencies has also been the standard solution, although multiresolution sinusoidal modeling was considered in [7]. Rate-distortion optimal time-segmentation [8] leads to an improved sinusoidal modeling, but still provides only a partial solution because a) the segmentation is still fixed over frequency and b) the minimum segment size is constrained because of the computational complexity involved in finding the optimum segmentation. Also, the use of overlap between segments inevitably smears any modeling error into neighboring frames.

In [9] amplitude modulated sinusoidal models for audio modeling and coding were introduced and in this paper we build further on this work. We achieve frequency-dependent temporal modeling using multiband amplitude modulation, where different amplitude modulating signals are used at different frequencies. Amplitude modulated sinusoidal models for audio modeling and coding are attractive for modeling of transient phenomena because constant-amplitude sinusoidal models converge slowly in terms of rate-distortion for transient signals thus performing badly for low bit-rates.

The paper is organized as follows. In Section 2 the amplitude modulated sinusoidal analysis-synthesis system is presented. This includes two parts, namely estimation of amplitude modulating signals and estimation of the parameters of the sinusoidal carriers. In Section 3 the multiband model is compared to the singleband model by listening tests as well as a perceptual distortion measure. Finally, conclusions about the work are presented in Section 4.

2 AM Sinusoidal Analysis-Synthesis

We use an amplitude modulated sinusoidal model, that looks as follows:

$$\hat{x}(n) = \sum_{q=1}^{Q} \gamma_q(n) \sum_{l=1}^{L_q} A_{l,q} \cos(\omega_{l,q} n + \phi_{l,q}), \tag{1}$$

where $\gamma_q(n)$ is the amplitude modulating signal in the q'th subband and L_q is the number of sinusoids in that subband. $\omega_{l,q}$, $A_{l,q}$ and $\phi_{l,q}$ are the frequencies, amplitudes and

phases of the sinusoids. We distinguish between a singleband model ($Q = 1$) and a multiband model ($Q > 1$).

The task is now to find $\gamma_q(n)$ for each subband. We start the estimation, which is based on [9], by splitting the input signal into subbands using the perfect reconstruction non-uniform filterbank described in [10]. Then we have for each subband a signal $x_q(n)$ and a model of that signal $\hat{x}_q(n)$. The instantaneous envelope of the model can then easily be shown to be

$$|\hat{x}_{q,c}(n)|^2 = \sum_{l=1}^{L_q}\sum_{k=1}^{L_q} \gamma_q^2(n) A_{l,q} A_{k,q} \times \exp(j(\phi_{k,q} - \phi_{l,q})) \exp(j(\omega_{k,q} - \omega_{l,q})n), \tag{2}$$

with subscript c denoting the analytic signal (see e.g. [9]). The squared instantaneous envelope is thus composed of a set of auto-terms ($l = k$) that identifies the amplitude modulating signal and a set of interfering cross-terms ($l \neq k$). From this it can be seen that the frequencies of these cross-terms in the instantaneous envelope are given by the distances between the sinusoidal components. Thus, the lowest frequency in the squared instantaneous envelope caused by the interaction of the sinusoids is given by the minimum distance between two adjacent sinusoids.

These cross-terms can be reduced by constraining the minimum distance between sinusoids and then lowpass filtering the squared instantaneous envelope of the input signal, i.e.

$$\gamma_q^2(n) = \alpha e_q^2(n) * h_{LP}(n), \tag{3}$$

where $e_q^2(n) = x_q^2(n) + \mathcal{H}\{x_q(n)\}^2$ with $\mathcal{H}\{\cdot\}$ denoting the Hilbert transform. Moreover, α is a positive scaling factor and $h_{LP}(n)$ is the impulse response of an appropriate lowpass filter with a stopband frequency below half the minimum distance between two sinusoids, i.e.

$$2BW < \min_{l \neq k} |\omega_{l,q} - \omega_{l,q}|. \tag{4}$$

For quasi-harmonic (pitched) sounds such as voiced speech, this spacing is simply the fundamental frequency. For a discussion on design issues regarding this filter see e.g. [5, 9].

Given that the amplitude modulating signal has been estimated for the q'th subband, we can then find the sinusoidal carriers of the subband (note that for convenience we now change the notation from indexing by subband to indexing by iteration). These are found by applying the estimated amplitude modulating signals to an overcomplete dictionary containing complex sinusoids resulting in a subband dictionary $\mathcal{D}_q$ containing atoms $g_{k,q}(n)$. We then perform matching pursuit [11], where in each iteration the maximizer of the normalized inner product between the atom and the residual is chosen, i.e.

$$\mathbf{g}_{i,q} = \arg\max_{g_{k,q} \in \mathcal{D}_q} \frac{| < \mathbf{g}_{k,q}, \mathbf{r}_{i,q} > |^2}{\|\mathbf{g}_{k,q}\|_2^2}, \tag{5}$$

where $\mathbf{g}_{k,q} = [g_{k,q}(0) \ \ldots \ g_{k,q}(N-1)]^T$ and $\mathbf{r}_{i,q} = [r_{i,q}(0) \ \ldots \ r_{i,q}(N-1)]^T$ with $r_{i,q}(n)$ being the residual of the i'th iteration. Now, writing out the inner product using the AM model, we get

$$< \mathbf{g}_{k,q}, \mathbf{r}_{i,q} >= \sum_{n=0}^{N-1} \gamma_q(n) \exp(-j\omega_k n) r_{i,q}(n). \tag{6}$$

It can be seen, that by defining $\tilde{r}_{i,q}(n) = \gamma_q(n) r_{i,q}(n)$, the greedy estimation can be carried out efficiently using an FFT of $\tilde{r}_{i,q}(n)$. In a similar way, we can apply the window $w(n)$ twice to the input and find the solution using an FFT, whereby the error is minimized in a weighted least-squares sense, i.e.

$$\begin{aligned} &\min \sum_{n=0}^{N-1} w^2(n) r_{i+1}^2(n) \\ =&\min \sum_{n=0}^{N-1} (w(n) c_k g_{k,q}(n) - w(n) r_{i,q}(n))^2, \end{aligned} \tag{7}$$

where c_k is the coefficient (phase and amplitude in this case) of the k'th atom. This causes not only the input but also the model to be weighted. This takes the use of windowing in both analysis and synthesis into account.

Equation (5) minimizes only the subband residual. When minimizing over the entire signal, we simply pick the maximum of the spectral subband maxima. This leads to the iterative (i being the iteration index) FFT-based algorithm (the FFT is denoted $\mathcal{FFT}\{\cdot\}$) described below, where the frequencies, phases and amplitudes of the sinusoidal model are found. We initialize the residuals with $r_{1,q}(n) = x_q(n) \ \forall q$.

1. Find subband

   $$q_i = \arg\max_q \left(\frac{|\mathcal{FFT}\{\gamma_q(n) w^2(n) r_{i,q}(n)\}|^2}{\sum_{n=0}^{N-1} \gamma_{q_i}^2(n) w^2(n)} \right)$$

   and corresponding frequency

   $$\omega_i = \arg\max_\omega |\mathcal{FFT}\{\gamma_{q_i}(n) w^2(n) r_{i,q_i}(n)\}|^2.$$

2. Estimate phase and amplitude by the inner product:

   $$c_i = \frac{\sum_{n=0}^{N-1} r_{q_i,i}(n) w^2(n) \gamma_{q_i}(n) \exp(-j\omega_i n)}{\sum_{n=0}^{N-1} \gamma_{q_i}^2(n) w^2(n)},$$

   which can be found from the subband FFT.

3. Generate new subband residual:

$$r_{i+1,q_i}(n) = r_{i,q_i}(n) - 2\gamma_{q_i}(n)|c_i|\cos(\omega_i n + \angle c_i).$$

This procedure is continued until some stopping criterion is reached. Although the estimation procedure is dependent on the amplitude modulating signal $\gamma_q(n)$, the algorithm still converges if we restrict $\gamma_q(n)$ to be strictly positive. Hereby the subband dictionaries $\mathcal{D}_q$ still form overcomplete bases and the algorithm converges on a subband level [11] and because of the perfect reconstruction filterbank, the entire system converges.

The above algorithm can be implemented much more efficiently than in the form above. The FFTs of the individual subbands and their maxima can be computed once at initialization. Then, in each iteration we only need to update the FFT of the subband residual and find the spectral maximum of it. The search in step 1 then reduces to searching among the Q spectral maxima.

3 Experimental Results

The importance of multiband temporal modeling has been investigated using both listening tests in the form of AB preference tests as well as an objective distortion measure. We compare the singleband model ($Q = 1$) to the multiband model ($Q > 1$).

Settings			
	Value		
Parameter	ABBA	GLCK	SPCH
Sampl. freq. [kHz]	44.1	44.1	8
Filterbank order	200	200	200
Filters	25	12	5
LP Filter order	100	100	100
Sinusoids	40	40	40
Cutoff freq. [Hz]	100	500	25

Table 1: Parameter values for different excerpts.

The excerpts used in the tests are: glockenspiel (GLCK), ABBA (ABBA), and Danish female speech (SPCH). They are all mono signals and have a length in the range of 5-10 s. These represent very different signal types from single source signals to complex music containing multiple sources.

The settings of the sinusoidal analysis-synthesis system for the different excerpts are shown in Table 1. In all cases a segment size of 20 ms and overlap-add with a 50% overlap von Hann window was used. Also, the FFT size was 8192. For the demodulation filter (3), we use an FIR filter designed using the window method (Hamming window).

In Table 2 the results of the AB preference tests are listed for the individual excerpts. 9 experienced listeners were used. It can be seen clearly, that there is a strong preference for the multiband model in the two cases, where the signals contain several sources, namely ABBA and glockenspiel, whereas for the case of speech, the preference tends toward equal. Significance has been determined by a small-sample case sign test (binomial distribution) using a 0.05 level of significance.

Results of Listening Tests			
	Preference		
Excerpt	Singleband	Multiband	Significant
ABBA	11%	89%	Yes
GLCK	11%	89%	Yes
SPCH	56%	44%	No

Table 2: Results of AB-preference tests.

Also, the results were verified using an objective measure. As a suitable perceptual model that also includes temporal masking phenomena, we used the Dau et al. model [12]. This model consists of a filterbank that resembles critical band filtering, followed by an inner-haircell model and adaptation loops which account for the temporal masking that occurs in the auditory system. The resulting internal representation is low-pass filtered and used for a perceptual distortion prediction by calculating the mean squared difference between the internal representations of the original and modified signal. The distortions are listed in Table 3. The Dau et al. model confirmed the results of the listening tests with the multiband model outperforming the singleband model in two first cases while the difference for the speech is very small. The overall distortion is highest for the ABBA excerpt, because it is a complex signal, whereas the total distortion is lowest for the speech signal, due to its limited bandwidth. That there is a slightly higher distortion for the multiband model for the case of SPCH can be attributed to the additional processing of the multiband system and the shape of the filters of the filterbank.

The conclusion is that for particular single sources such as speech, the singleband model performs very well. This is in line with [5], where also members of brass, wood-

Results using the Dau et al. Model		
	Distortion	
Excerpt	Singleband	Multiband
ABBA	7270	5431
GLCK	931	644
SPCH	412	426

Table 3: Distortions calculated using the Dau et al. model.

wind, and string instrument families are mentioned as sources being well modeled by the singleband model. For more complex signals such as superpositions of multiple sources, there is a great need for multiband modeling and coding, which is clearly indicated by the high preference for multiband modeling of ABBA.

That the singleband model works well for single sources is an indication that the model in Eq. (1) can indeed form the basis of compression not only in terms of subbands, but also in terms of sinusoids sharing an amplitude modulating signal, i.e. by a decomposition into sources.

4 Conclusion

In this paper, we have investigated the need for taking temporal phenomena in audio modeling and coding into account in a way that is frequency dependent. This has been done in the context of sinusoidal modeling, where we have applied amplitude modulation in order to achieve better temporal modeling. We have presented a multiband sinusoidal analysis-synthesis system that utilizes amplitude modulation to achieve frequency dependent temporal resolution. Finally, we have compared this multiband model to a commonly used singleband model and it has been demonstrated using both an objective perceptual distortion measure as well as listening tests, that significant improvements are achieved by this for complex signals containing multiple sources such as general audio, and that the singleband model performs very well for particular single sources.

References

[1] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88(4), pp. 451–515, Apr. 2000.

[2] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, 2nd ed. Springer, 1999.

[3] B. C. J. Moore, "Masking in the human auditory system," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. Audio Eng. Soc., 1996, pp. 9–19.

[4] B. Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," *Frequenz*, pp. 1033–1036, 1989.

[5] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40(6), pp. 497–516, 1992.

[6] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4.

[7] S. N. Levine, T. S. Verma, and J. O. Smith III, "Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 1997, pp. 101–104.

[8] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[9] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal models for audio modeling and coding," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Artificial Intelligence, V. Palade, R. J. Howlett, and L. C. Jain, Eds. Springer-Verlag, 2003, vol. 2773, pp. 1334–1342.

[10] M. M. Goodwin, "Nonuniform filterbank design for audio signal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1997, pp. 1229–1233.

[11] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[12] 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. Am.*, vol. 99(6), pp. 3615–3622, June 1996.

Paper C

Efficient Parametric Coding of Transients

Mads Græsbøll Christensen and Steven van de Par

The paper will appear in
IEEE Transactions on Audio, Speech and Language Processing, vol. 14(4),
July 2006.

The layout has been revised.

Abstract

In this paper, methods for improved parametric coding of transients are presented. We propose a signal model for coding of transients consisting of a sum of sinusoids each being amplitude-modulated by a different gamma envelope. These envelopes are characterized by an onset time, an attack and a decay parameter. An efficient method for estimating these parameters is presented. Further, methods are proposed that combine this transient model with a constant-amplitude sinusoidal model in order to achieve efficient coding of both stationary and transient signal parts. By rate-distortion optimization using a perceptual distortion measure we combine variable rate bit allocation and segmentation in an optimal way. Formal as well as informal listening tests show that significant improvements can be achieved with the proposed model as compared to a state-of-the-art sinusoidal coder by the combination of optimal segmentation and amplitude modulated sinusoidal audio coding.

1 Introduction

In the past couple of decades, sinusoidal models for digital processing of speech and audio have received much attention for a wide variety of applications where sinusoidal speech coding and modeling [1–4] was among the first and perhaps the most prominent. Also for analysis and synthesis of music [5, 6] the sinusoidal model has been of interest. In recent years, the growth of the Internet and wireless communication has spurred renewed interest in sinusoidal models, this time for coding of audio [7–15] at low bit-rates. In perceptual audio coding, compression is achieved by exploiting statistical redundancies as well as perceptual irrelevancies of the source (see e.g. [16]). In parametric audio coding, a compact representation of the source signal is achieved using parametric models and the statistical redundancies and irrelevancies of the model parameters are exploited for efficient coding.

A major challenge in audio coding in general is efficient coding of non-stationary segments (see e.g. [16]). Signal models and transform bases are typically chosen such that a high coding efficiency is achieved for stationary signal parts, and, as a consequence, coding of non-stationary parts becomes highly inefficient. Sinusoidal coding using constant-amplitude (CA) sinusoids is an example of this difficulty. The inefficient coding of transients leads to a number of problems. Firstly, errors introduced before onsets are very poorly masked compared to the situation where a simultaneous masker is present [17]. These types of errors are known as pre-echos. Secondly, bad modeling of transients leads to very dull sounding attacks and a perceived lack of bandwidth of the decoded signal. The typical solution to these problems are adaptive segmentation using window switching [18] and window shape adaptation or rate-distortion (R-D) optimal segmentation [14, 19, 20]. Other methods that aim at solving this problem include wavelet-packets [21], temporal noise shaping (TNS) [22], gain modification [23, 24],

transient location modification [25], switching from a parametric signal model to a wavelet or transform representation [7, 9], multi-resolution sinusoidal modeling [26] and coding of transients using sinusoidal modeling in the transform domain [27]. In parametric audio modeling and coding, transients can be handled by adapting the signal model to better fit the input signal. A particularly interesting class of such adapted models are the amplitude modulated (AM) sinusoidal models[1] [28]. In these models, the signal is decomposed into a sum of sinusoidal components having a time-varying envelope. The different realizations of damped sinusoids that have been applied to audio modeling in [29–33] are examples of this. In audio coding AM has been applied in [8, 13]. Like [5] these use a singlebanded model of the modulating signal meaning that the envelope is the same for all components. In [34] it was demonstrated that significant improvements are achieved by allowing different sinusoidal components to have different amplitude modulating signals. Since this study focused only on modeling of audio signals, the question remains whether frequency-dependent AM methods are also efficient in terms of bit-rate, i.e., whether they achieve a lower distortion, both subjectively and objectively, compared to a conventional sinusoidal coder at the same rate.

In the present paper we seek to answer that question along with some other unanswered questions regarding parametric coding of transients. We present a coder based on a particular model of the amplitude modulating signal known as gamma envelopes. Figure 1 shows the waveform of a sinusoid modulated by a windowed gamma envelope. The gamma envelopes are characterized by an onset time, an attack and a decay parameter. This model differs from existing models used for parametric modeling and coding of audio in that each sinusoid can have a different envelope with an onset at an arbitrary position within a segment, and in that it is characterized by an attack parameter. In addition to the new signal model, the proposed coder incorporates rate-distortion optimal bit allocation and segmentation. Further, we consider different ways of achieving efficient coding of both stationary and transient signal parts. Finally, we quantify, by subjective listening tests, the performance of the different methods for different types of signals.

The main part of this paper is organized as follows: in Section 2 the proposed signal model and the perceptual distortion measure which is instrumental in this work are presented. The rate-distortion optimization used for allocation and segmentation is presented in Section 3, and Sections 4 and 5 deal with the estimation of sinusoidal parameters. Implementation details, the experimental setup for perceptual tests and their results are presented in Sections 6 and 7, respectively. In Section 8 we discuss the relation to existing work, and, finally, in Section 9 we conclude on our work.

[1] In this text, AM means either amplitude modulation or amplitude modulated depending on the context.

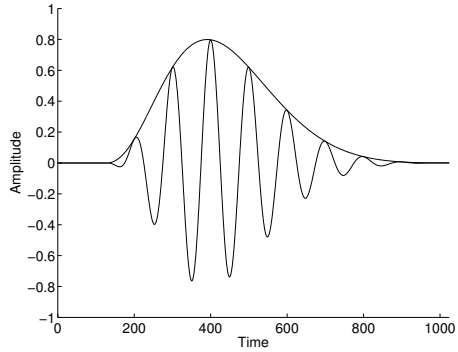

Figure 1: Illustration of a sinusoid modulated by a windowed gamma envelope. The gamma envelopes are parameterized by an onset, an attack parameter and a decay parameter.

2 Fundamentals

The presented coder can be described as comprising the following steps: in the encoder, the input signal is split into a number of overlapping segments and a window is applied to each segment. The model parameters are then estimated and subsequently quantized, entropy coded and finally put into the bit-stream. In the decoder, the bit-stream is mapped back to the quantized parameters, and the segment is synthesized using overlap-add with an appropriate window.

In this paper, we propose a coder based on the following amplitude modulated sinusoidal signal model for time index $n = 0, \ldots, N-1$:

$$\hat{x}(n) = \sum_{l=1}^{L} \gamma_l(n) A_l \cos(\omega_l n + \phi_l), \tag{1}$$

where A_l, ω_l, and ϕ_l are the amplitude, frequency and phase of the l'th sinusoids, respectively. The number of components is denoted L and $\gamma_l(n)$ is the modulating signal or envelope when $\gamma_l(n) \geq 0 \; \forall n$. Here we use a particular model of the envelopes which we shall henceforth refer to as gamma envelopes. This model is derived from the integrand of the gamma function, which is commonly used to characterize the gamma distribution in statistics. The gamma envelopes are given as

$$\gamma_l(n) = u(n - n_l)\,(n - n_l)^{\alpha_l}\, e^{-\beta_l (n - n_l)}. \tag{2}$$

Each envelope is characterized by an onset time $n_l \in \mathbb{Z}$, an attack parameter $\alpha_l \in \mathbb{N}$, and a decay parameter $\beta_l \in \mathbb{R}^+$. Moreover, $u(n)$ is the unit step sequence. The envelopes composed from all possible combinations of these parameters will henceforth be referred to as the envelope dictionary. Inserting (2) into (1), we get the so-called

gamma-tones commonly used as stimuli in psychoacoustical experiments and for modeling of the auditory filters [35]. Here, we rather use it as a signal model that, as we shall see, has been found to perform well for the problem at hand. The distinction between the model parameters α_l and β_l in (2) is only figurative since changing β_l for a fixed α_l will affect the attack and α_l will likewise affect the decay. We note that for $\alpha_l = 0$, $\beta_l = 0$ and $n_l = 0$, the lth sinusoid reduces to a constant-amplitude (CA) sinusoid, i.e. $\gamma_l(n) = 1$. The situation where all components have constant amplitude will be termed the CA model. For $\alpha_l = 0$ and $\beta_l \neq 0$ for all l, the model reduces to the so-called delayed damped sinusoids of [32], and with $\alpha_l = 0$ and $n_l = 0$ it becomes equivalent to the damped sinusoids of [30, 33]. Compared to the different variations of damped sinusoids of [29–32], this model has the additional flexibility of the attack parameter. It is well-known that different instruments do have different attacks, and studies show that the attacks are in fact important features in the recognition of musical instruments [36]. This can also be witnessed from the many transient signals on the SQAM disc [37].

In finding the model parameters and in the R-D optimization, it is advantageous to use a perceptual distortion measure since we seek to minimize the perceived distortion. In choosing a distortion measure we face conflicting demands. On one hand we wish to use a distortion measure that takes as much of the human auditory system into account as possible. On the other hand we wish to have a distortion measure that is both of reasonably low computational complexity and defines a norm such that it may be subject to optimization. Consequently, we have chosen the spectral distortion measure of [38], which is defined as

$$D = \int_{-\pi}^{\pi} A(\omega)|E(\omega)|^2 d\omega, \tag{3}$$

where $A(\omega)$ is a real, positive perceptual weighting function, and $E(\omega)$ denotes the discrete-time Fourier transform of the windowed error, i.e.,

$$E(\omega) = \sum_{n=0}^{N-1} w(n)e(n)e^{-j\omega n}, \tag{4}$$

with $w(n)$ being the analysis window, $e(n) = x(n) - \hat{x}(n)$ the modeling error, and $x(n)$ the observed signal. We note in passing that this and all other Fourier transforms will in practice be calculated for discrete values of ω. In order to shape the error spectrum according to the masking threshold, the weighting function $A(\omega)$ is set to the reciprocal of the masking threshold. Here, we derive the masking threshold from [38]. This distortion measure improves on other models in that it takes the spectral integration in the human auditory system into account. Although the measure is strictly only valid for stationary signals, it does not ignore temporal aspects completely as it is based on waveform matching. In order to achieve a low distortion, the phase and temporal envelope of the coded signal must match that of the original. As a consequence, temporal

errors, such as pre-echos, will not go unpunished by the measure. The spectral distortion measure has been found to comprise a reasonable tradeoff between complexity and correlation with perceived quality for coding purposes and as we shall see, good results can be achieved using it. Henceforth, when we refer to distortions, we mean the perceptual distortion defined in (3).

The discrete-time Fourier transform of $\gamma_l(n)$ denoted $\Gamma_l(\omega)$ can be shown to be

$$\Gamma_l(\omega) = \sum_{n=0}^{N-1-n_l} n^{\alpha_l} e^{-j\omega n_l} \left(e^{-j\omega-\beta_l}\right)^n \tag{5}$$

$$= j^{\alpha_l} \frac{\partial^{\alpha_l}}{\partial \omega^{\alpha_l}} \frac{e^{-j\omega n_l} - e^{-\beta_l(N-n_l)} e^{-j\omega N}}{1 - e^{-\beta_l} e^{-j\omega}}. \tag{6}$$

As indicated by (4), an analysis window is applied to the gamma envelopes. In the decoder, a window is also used in the synthesis, which is performed using overlap-add with a fixed overlap. Both the encoder and the decoder use tapered von Hann windows of the same length. With M denoting the overlap in samples and N being the (even) segment length, the windows are defined for $n = 0, \ldots, N-1$ as

$$w(n) = \begin{cases} v(n), & 0 & \leq n < & M \\ 1, & M & \leq n < & N-M \\ v(n-N+2M), & N-M & \leq n < & N \end{cases} \tag{7}$$

with the even length von Hann window being defined as

$$v(n) = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{\pi(n+0.5)}{M}\right). \tag{8}$$

Let $W(\omega)$ denote the discrete-time Fourier transform of the window $w(n)$. Then the discrete-time Fourier transform of the windowed envelope can be written as the circular convolution

$$Z_l(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Gamma_l(\omega - \xi) W(\xi) d\xi. \tag{9}$$

Hence, the window, which has low-pass characteristics, smoothes the spectrum. As the windowed gamma envelopes have no discontinuities at segment boundaries the spectrum of the windowed gamma envelopes will generally be more well-behaved than when no window is applied. This is important since the distortion measure will punish spectral distortion due to not only the mainlobe but also the sidelobes. In Appendix A, a closed-form expression of the discrete-time Fourier transform of the windowed gamma envelopes is derived.

3 R-D Optimal Allocation and Segmentation

Since audio signals may exhibit varying degrees of stationarity, it is often advantageous to allow for a flexible segmentation and allow the bit-rate to vary over time. In addition,

it is observed that the proposed AM signal model is only efficient in terms of rate-distortion for transient segments, while the CA model is an efficient representation of tonal stationary segments. In order to combine the two models in an optimal way as well as doing optimal segmentation of the input signal, we use rate-distortion optimization. Further, the rate-distortion optimization also results in a rate-scalable coder, which is advantageous in dealing with critical signal parts. For completeness we now briefly review the basic definitions, assumptions and results for solving the problem of optimal segmentation and allocation based on [19, 39]. First, let us start out by introducing some definitions. We define a segment σ_s as having a length of a positive integer multiple $m \in \mathbb{Z}^+$ of a minimum segment length κ, i.e. $\ell(\sigma_s) = \kappa m$, and a segmentation as $\boldsymbol{\sigma} = [\, \sigma_1 \ \cdots \ \sigma_S \,]$ consisting of S disjoint, contiguous segments that satisfy

$$\sum_{s=1}^{S} \ell(\sigma_s) = \kappa M, \tag{10}$$

where κM is the total length of the signal to be encoded. Each of these segments, say segment σ_s, can then be encoded using a set of coding templates $\mathcal{T}_s$ (different models, model orders, number of bits, etc.). Next, we define $R(\sigma_s, \tau_s)$ and $D(\sigma_s, \tau_s)$ as the non-negative cost in bits and distortion associated with coding template $\tau_s \in \mathcal{T}_s$ for segment σ_s. Assuming that the distortions and cost in bits associated with a particular segmentation $\boldsymbol{\sigma}$ and coding templates $\boldsymbol{\tau} = [\, \tau_1 \ \cdots \ \tau_S \,]$ are additive over the segments, we can write the total distortion and total number of bits as

$$D(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \sum_{s=1}^{S} D(\sigma_s, \tau_s) \quad R(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \sum_{s=1}^{S} R(\sigma_s, \tau_s), \tag{11}$$

respectively. The problem of distributing a certain number of bits over a number of quantizers can be cast into the problem of rate-distortion optimization under rate constraint. This can be stated as the following constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & D(\boldsymbol{\sigma}, \boldsymbol{\tau}) \\ \text{s. t.} \quad & R(\boldsymbol{\sigma}, \boldsymbol{\tau}) \leq R^\star, \end{aligned} \tag{12}$$

with $R^\star$ being the bit budget, i.e. the total number of bits to be distributed. Next, introducing the Lagrange multiplier $\lambda \geq 0$, the constrained optimization problem in (12) can be written as the unconstrained minimization problem [39]

$$J(\lambda) = \min_{\boldsymbol{\sigma}} \min_{\boldsymbol{\tau}} \sum_{s=1}^{S} D(\sigma_s, \tau_s) + \lambda (R(\sigma_s, \tau_s) - R^\star). \tag{13}$$

We now have an outer minimization over the segmentation, and an inner minimization over coding templates given the segmentation. In (11) we assumed that $D(\cdot)$ and

$R(\cdot)$ are additive over segments. By also assuming that they are independent over segments, the inner minimization in (13) can be simplified significantly. Specifically, the optimization problem reduces to the following, where the coding templates can be optimized independently for a segmentation and a particular λ [19]:

$$J(\lambda) = \min_{\boldsymbol{\sigma}} \sum_{s=1}^{S} \min_{\tau \in \mathcal{T}_s} \left[D(\sigma_s, \tau) + \lambda R(\sigma_s, \tau) \right] - \lambda R^\star. \tag{14}$$

This leads to the following important result: as the rates and distortions are additive over segments, the outer minimization can be solved using dynamic programming [19]. The optimal λ that leads to the target rate $R^\star$, denoted $\lambda^\star$, can be found by maximizing the concave Lagrange dual function [40], i.e.,

$$\lambda^\star = \arg\max_{\lambda} J(\lambda) \tag{15}$$

This can be done by sweeping over λ until $R(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is within some range of the bit budget [19]. It should be noted that for a discrete problem such as ours, we cannot guarantee that strong duality holds for the optimization problem, and, as a consequence, the found solution may be suboptimal, but for a dense set of coding templates the gap will be small (see [40]). For a fixed segmentation, i.e. given $\boldsymbol{\sigma}$, the outer minimization disappears, and we only have to minimize over the coding templates. This was the approach used in [41].

4 Parameter Estimation

The distortion measure (3) defines a norm and is in fact induced by an inner product (see [42]). The parameters for each sinusoid can then be found using a matching pursuit algorithm [43]. This would guarantee convergence in the distortion as a function of the number of components. The psychoacoustic matching pursuit (PMP) [42] is an algorithm that does this, i.e. it performs matching pursuit using the norm (3). The inner products can be found using FFTs also for the AM case. It would, however, be very expensive with respect to computational complexity. Since the R-D optimal segmentation requires that at every segment boundary, all combinations of segment lengths and coding templates are evaluated, it is critical that the estimation procedure is fast. In that spirit, we here employ a simpler procedure than PMP. We start out by noting the number of different combinations of parameters will be dominated by the number of different frequencies and onset points. Thus, we break the estimation process into three successive steps: frequency estimation, onset estimation, and, finally, estimation of the envelope parameters and the corresponding phase and amplitude. A block diagram of the estimation procedure is shown in Figure 2.

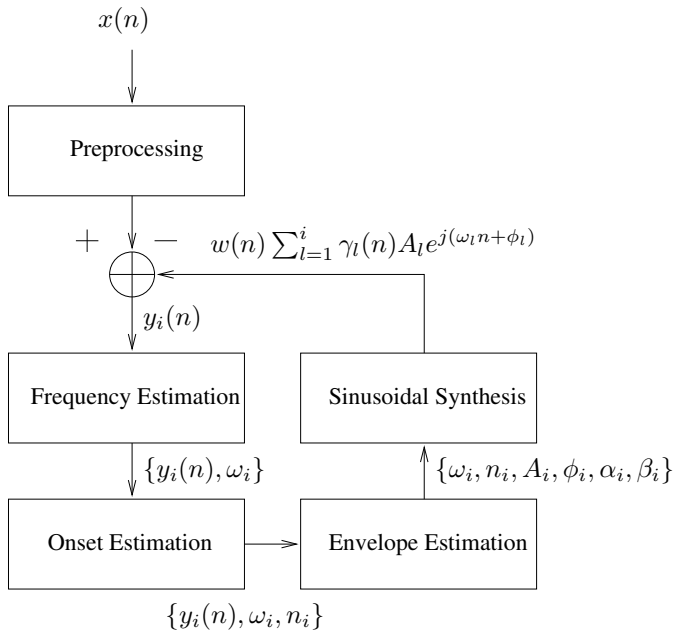

Figure 2: The iterative AM parameter estimation procedure. Sinusoids are found one at the time and subtracted from the input.

For the frequency estimation we use a fast method somewhat reminiscent of the weighted matching pursuit [44]. The algorithm operates on the residual, which at iteration $i+1$ is formed as

$$y_{i+1}(n) = y_i(n) - w(n)\gamma_i(n)A_i e^{j(\omega_i n+\phi_i)}. \tag{16}$$

The residual is initialized as the discrete-time analytic signal

$$y_1(n) = w(n)x(n) + jw(n)\mathcal{H}\left\{x(n)\right\}, \tag{17}$$

where $\mathcal{H}\left\{\cdot\right\}$ denotes the Hilbert transform. This, including windowing, is the preprocessing step in Figure 2. In practice, the Hilbert transform is found using the FFT method. By operating on the analytic signal, we ignore the spectral contents of $x(n)$ for negative frequencies. This is done in order to simplify the estimation procedure. Convergence in the modeling of the analytic signal also ensures convergence in the real signal since

$$\Re\left\{w(n)x(n) + jw(n)\mathcal{H}\left\{x(n)\right\}\right\} = w(n)x(n), \tag{18}$$

however, for a non-zero error, the analytic signal modeling will introduce some error due to the correlation between negative and positive sides of the spectrum.

Let $P_i(\omega) = Y_i^*(\omega)Y_i(\omega)$ be the squared magnitude of the discrete-time Fourier transform of the residual at iteration i , i.e.,

$$Y_i(\omega) = \sum_{n=0}^{N-1} y_i(n)e^{-j\omega n}, \tag{19}$$

which may be updated efficiently in the frequency domain. Then the frequency is estimated as

$$\begin{aligned} \omega_i &= \arg\max_{\omega} A(\omega)P_i(\omega) \\ \text{s. t.} \quad & \frac{\partial P_i(\omega)}{\partial \omega} = 0 \quad \text{and} \quad \frac{\partial^2 P_i(\omega)}{\partial \omega^2} < 0. \end{aligned} \tag{20}$$

This estimation criterion can be seen as an asymptotic PMP criterion with $N \to \infty$ for the CA case. The constraints ensure that the frequency will be a peak in the spectrum. This is a reasonable restriction also for the AM case as the modulating signals all have low-pass characteristics. We cannot, however, guarantee that the error converges in a convex way.

A coarse estimate of the integer onset n_i is found in order to limit the search space using the following simple method: given a model where a sinusoidal component of frequency ω_i is modulated by a unit step sequence $u(n-\zeta)$, the modeling error can be written as

$$y_i(n) - w(n)u(n-\zeta)A_i e^{j(\omega_i n + \phi_i)}. \tag{21}$$

This error is minimized in a least-squares sense by maximizing the inner product (with proper normalization) between the modulated sinusoid and the residual:

$$\Psi(\zeta) = \frac{1}{\sum_{n=\zeta}^{N-1} w^2(n)} \left| \sum_{n=\zeta}^{N-1} y_i(n)w(n)e^{-j\omega_i n} \right|^2. \tag{22}$$

We note that the product $y_i(n)w(n)e^{-j\omega_i n}$ for $n = 0, \ldots, N-1$ only has to be computed once for each sinusoid. We then find the onset as the maximizer of (22), i.e.,

$$n_i = \arg\max_{\zeta} \Psi(\zeta). \tag{23}$$

Given the frequency and the coarse onset, the combination of envelope parameters, including a final onset estimate, is found as the minimizer of the distortion measure (3). This corresponds to performing a PMP on the subset of the dictionary. We assume that all the dictionary elements have been scaled for a particular segment such that they all have unit perceptual norm, i.e.,

$$\int_{-\pi}^{\pi} A(\omega)Z_k^*(\omega-\omega_i)Z_k(\omega-\omega_i)d\omega = 1 \quad \forall k, \tag{24}$$

with Z_k being the discrete-time Fourier transform of the windowed envelope k in the dictionary, i.e. (see Appendix A)

$$Z_k(\omega) = \sum_{n=0}^{N-1} w(n)\gamma_k(n)e^{-j\omega n}. \tag{25}$$

The envelope, i.e. the combination of α_i, β_i and n_i, is then found in an analysis-by-synthesis manner as the minimizer of the perceptual distortion or, equivalently, as the following maximization of the inner product:

$$Z_i(\omega) = \arg\max_{Z_k(\omega)} \left| \int_{-\pi}^{\pi} A(\omega) Z_k^*(\omega - \omega_i) Y_i(\omega) d\omega \right|^2. \tag{26}$$

From this inner product, the phase and amplitude of the i'th sinusoid can also be found as the modulus and the argument, i.e.

$$A_i e^{j\phi_i} = \int_{-\pi}^{\pi} A(\omega) Z_i^*(\omega - \omega_i) Y_i(\omega) d\omega. \tag{27}$$

In practice the spectra are discrete and the integration is performed as a summation over point-wise multiplications. As most of the spectral energy of $Z_i(\omega - \omega_i)$ is concentrated in a small region around ω_i, the integration range can also be reduced without much loss in accuracy but with considerable reduction of computational complexity.

For the segment lengths used here, the analytic signal model (considering only the positive parts of the spectrum) has been found to perform satisfactorily. We note that it is also possible to account to some extent for the interaction between different components, including the positive and negative sides of the spectrum, in a number of different ways. The different well-known optimizations of matching pursuit (see e.g. [45]) can be applied at the cost of additional complexity since (3) defines a norm.

5 Rate-Regularized Estimation

In section 4, the parameter set of each envelope, denoted $\Omega_i = \{ \alpha_i \; \beta_i \; n_i \}$, was found in iteration i as the minimizer of the distortion

$$\hat{\Omega}_i = \arg\min_{\Omega_i} D(\Omega_i), \tag{28}$$

or equivalently as the maximization in (26). Since sinusoids having constant amplitude do not require the envelope parameters to be transmitted, disregarding the rate in the estimation results in a parameter set which is suboptimal in a rate-distortion sense. In [41] every segment was analyzed using a set of constant-amplitude sinusoids and a set of amplitude modulated sinusoids and by rate-distortion optimization the best representation

was chosen for each segment. This was done in order to find an efficient representation in terms of rate. Suppose we have an estimate, or a guess, of $\lambda^\star$ denoted ν, the need for multiple analyses can be eliminated by instead minimizing in each iteration of the estimation

$$\hat{\Omega}_i = \arg\min_{\Omega_i} \left[D(\Omega_i) + \nu R(\Omega_i) \right], \tag{29}$$

where $R(\Omega_i)$ denotes the rate associated with the parameters Ω_i. The rate-distortion optimization is still performed outside the estimation such that the rate-constraint is met. The rate-regularized estimation procedure results in coding templates that are optimized for the target bit-rate. As an example, consider the choice in iteration i between an amplitude modulated sinusoid and a constant-amplitude sinusoid. Using the estimation criterion in (28), the amplitude modulated sinusoid may be chosen, while using (29) may result in the constant-amplitude sinusoid being chosen because the amplitude modulated sinusoid is more expensive in terms of rate. The estimation criterion (29), which we from now on shall refer to as the rate-regularized estimation or just regularized estimation, corresponds to optimizing the coding templates for the target bit-rate. The regularization constant ν does not, however, play the role of the Lagrange multiplier in constrained optimization since we do not solve for it. By choosing $\nu = 0$, the estimation criterion will reduce to (28). Using a large ν will result in an estimation that will tend to choose constant-amplitude over amplitude-modulated sinusoids, while for a small ν, the opposite will occur. In the extremes, this will result in a coder containing only constant-amplitude or amplitude modulated sinusoids. It must be stressed that even if $\nu = \lambda^\star$, i.e. if we guessed the optimal ν, the estimation is not optimal as the individual iterations are not independent. It is of course possible to iterate over ν, but this would be costly in terms of complexity. In most practical situations, the actual choice of ν has been found not to be very critical, i.e., it can simply set to a constant value.

6 Implementation Details

6.1 Sinusoidal Parameter Quantization and Rate Estimates

The phases of the sinusoidal components are quantized uniformly using 5 bits, while amplitudes and frequencies are quantized in the logarithmic domain using the following quantizers.With θ denoting the parameter to be quantized and $\lfloor \cdot \rfloor$ the truncation operation, the quantized parameter $\hat{\theta}$ is calculated as

$$\hat{\theta} = \exp\left(\left\lfloor \frac{\log(\theta + \epsilon)}{\log(1+\Delta)} + 0.5 \right\rfloor \log(1+\Delta) \right), \tag{30}$$

with a small positive constant ϵ being added for numerical reasons. With a step-size Δ of 0.161 for the amplitudes and 0.003 for the frequencies, the quantizers were found

to produce transparent results compared to the original (non-quantized) parameters, meaning that informal listening tests showed no degradation in the perceived quality due to the quantization. These quantizers are motivated by studies that show that for amplitude and frequency the just noticeable differences are nearly constant on a logarithmic scale [46]. Estimated entropies of the quantized parameter sets were used for the rates in the R-D optimization and as a measure of rate in the experiments to follow. The entropies of the quantized sinusoidal parameters were also found not to be affected much by the AM. For the amplitude, phase and frequency the entropy was estimated as approximately 20 bits/component. Assuming differential encoding [47], this can be reduced to 16 bits/component. Since the perceptual distortion measure (3) may be overly sensitive to frequency quantization, we use the original parameters in determining the distortions. For the same reason the original parameters are used in generating the residual in the estimation (16).

6.2 Coding Templates and Segment Sizes

In the experiments to follow, a number of different coder configurations were considered. These are listed in order of rising complexity in Table 1. The table shows what types of coding templates were used, how they were found and whether R-D optimal segmentation (SEG) was used. The coding templates are defined as $\mathcal{T}_s = \{\chi_0, \ldots, \chi_L\}$, where χ_i means i sinusoids, which may or may not be modulated, depending on the type of coder. For example, the AM/CA coder uses fixed segmentation and contains coding templates found by analyzing a particular segment using a set of AM sinusoids and a set of CA sinusoids. Note that the AM coding templates can contain constant-amplitude components since these are included as a special case of the model (2), while the CA coding templates contain only CA components. In order to efficiently code CA components in the AM coding templates, a one bit AM switch is used per component. This may be more efficiently encoded using run-length coding. The CA+SEG coder is comparable in quality to that of [48], which uses the PMP and R-D optimal segmentation and uses identical quantizers. The segmentation algorithm described in Section 3 requires that the distortions are additive over segments. For this to be true, the segments have to be disjoint. However, in order to avoid discontinuities at segment boundaries, some amount of overlap must be introduced between adjacent segments. That the errors introduced in the overlapping regions may have non-zero cross-terms is then simply ignored. Since the distortions also have to be independent over segments, the amount of overlap between segments cannot depend on the segment length. Therefore a natural choice for the amount of overlap is half the size of the minimum segment length. It is important that the overlap is not too small since this may cause undesirable artifacts due to quantization and estimation errors. Consequently, a minimum segment length of 10 ms and an overlap of 5 ms is chosen, meaning that all segment sizes are integer multiples of 10 ms and may start on a 5 ms time-grid. Further, for very long segments, the spectral weighting function becomes increasingly inaccurate as the maskers cannot

Coder	Description
CA	The CA coder uses coding templates consisting of constant-amplitude sinusoids only and a fixed segmentation. This is the simplest possible coder.
AM	The AM coder uses amplitude modulated coding templates and a fixed segmentation. This coder uses the rate-regularized estimation procedure using a regularization constant of 100.
AM/CA	A combination of the CA and AM coder operating on a fixed segmentation. It switches between the two on a segment-to-segment basis using R-D optimization. It does not use the rate-regularized estimation procedure, i.e. a regularization constant of 0 is used.
CA+SEG	As the CA coder but with R-D optimal segmentation.
AM+SEG	The same as the AM coder but with R-D optimal segmentation.
AM/CA+SEG	This is the AM/CA coder combined with R-D optimal segmentation.

Table 1: Coder configuration for different test cases denoted by coder acronym.

be assumed to be stationary. Therefore a maximum length of 40 ms has been used. For the coders that use a fixed segmentation, a von Hann window of 30 ms with 15 ms overlap was used. In the experiments to follow, we ignore the side information associated with the segmentation, as this can generally be considered small compared to the total rate. Moreover, the critical comparisons are between coders that use the same type of segmentation and thus have the same rate for the side information. The excerpts used in the tests to follow are fairly short, and the rate-distortion optimization has therefore been carried out over the entire length of the signals.

6.3 Gamma Envelope Dictionary

It has been found that using the perceptual distortion measure (3) in selecting the envelope parameters made the parameter estimation more robust toward introducing artifacts than using a squared error measure. This can be attributed to the fact that the spectral distortion measure takes into account that the wide mainlobe and sidelobes of modulated sinusoids may introduce errors in parts of the spectrum where no masker is present. However, it was also found necessary to limit the steepness of the attack in order to prevent artifacts from being introduced. Namely, we found that for small α_l, the coder was prone to introduce roughness and click artifacts due to the discontinu-

Number	Name	Type	Length
1	Castanets and Guitar	Mixed	6 s
2	Claves	Solo	7 s
3	Glockenspiel	Solo	8 s
4	Grand Piano	Solo	11 s
5	ABBA	Mixed	10 s
7	Bass Guitar	Solo	12 s
8	English Female Speech	Speech	6 s
9	Castanets	Solo	7 s
10	Harpsichord	Solo	9 s
11	Tracy Chapman	Mixed	13 s
12	Triangle	Solo	9 s
13	Xylophone	Solo	8 s

Table 2: List of excerpts used in the tests.

ities introduced by the unit step sequence. We again note that for $\alpha_l = 0$, the model reduces to that of [32]. Hence, the envelope dictionary was designed empirically from the results of informal listening tests. With a more refined distortion measure, the envelope dictionary could be designed using standard vector quantization techniques. In the following tests, an envelope dictionary for a sampling frequency of 48 kHz composed from $\alpha_l \in \{2, 3, 4, 5\}$, $\beta_l \in \{0.003, 0.005, 0.01, 0.02\}$ and an onset n_l step-size of approximately 0.5 ms was used. As a consequence of this the envelope dictionary size varies with the segment lengths. Since the frequency and envelopes of transients may vary much from signal to signal, no entropy coding of the envelope parameters was assumed in the rate estimates, i.e. the upper bound is used. These are 9, 10, 10 and 11 bits per envelope for 10, 20, 30 and 40 ms segments, respectively. Preliminary experimental results also suggest that differential coding of onset times may lead to a reduction of the average bits per component. The spectra of the windowed gamma envelopes were stored in a lookup table in order to perform fast estimation (equations (26) and (27)) using the spectral distortion measure (3).

7 Experimental Results

7.1 Signal Examples

As an example of a coded signal, the xylophone coded at 30 kbps is shown in Figures 3 and 4. It can be seen that the CA coder introduces a pre-echo and that the transient is smeared and has lost its sharpness. In the CA+SEG coder, the pre-echo is much reduced, but the transient is still not as sharp as the original. The AM/CA+SEG coder

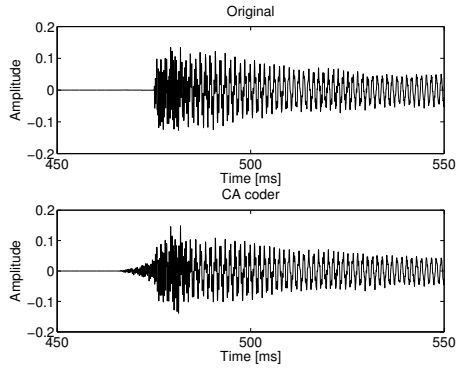

Figure 3: Signal example, xylophone, original (top) and coded at 30 kbps using the CA coder (bottom).

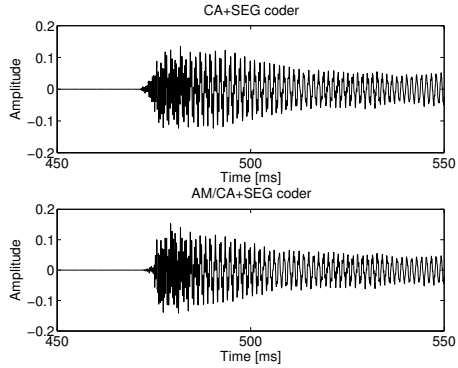

Figure 4: Signal example, xylophone, coded at 30 kbps using the CA+SEG coder (top) and using the AM/CA+SEG coder (bottom).

sharpens the attack further and reduces the pre-echo.

In Figure 5 the rate-distortion curves[2] for a representative transient sinusoidal signal, glockenspiel, are shown for the CA coder, the AM/CA coder and the AM coder. Similarly, in Figure 6, the same is shown for the CA+SEG coder, the AM/CA+SEG coder and the AM+SEG coder. The signal has a duration of approximately 10 s and R-D optimization was performed on the entire signal. For the fixed segmentation, it can be seen that there is a clear improvement for the AM and AM/CA coders in terms of a reduction of the distortion compared to the CA coder at the same rate. Also, the proposed coder saturates at lower distortions than the CA coder for glockenspiel. It can also be seen from Figure 6, that when R-D optimal segmentation is employed, the rate

[2]In information theory the relation $D(R)$ is traditionally referred to as the distortion-rate curve. We refer to this relationship using the aesthetically more pleasing term rate-distortion curve.

of convergence is higher for all coders. An interesting observation is also that the rate-regularized coder, the AM coder, performs similarly to the AM/CA coder. This means that the dual analyses of the AM/CA coder can be avoided with very little loss of performance. From these figures, it seems that for this particular excerpts, the glockenspiel, very little is achieved by combining AM and SEG. It looks as if similar performance can be achieved with either AM or SEG, with the AM coder being less complex than the CA+SEG coder. For other signals such as the castanets, though, the R-D curves show that improvements can be gained by the combination of AM and R-D optimal segmentation.

In Figure 7 the R-D optimal segmentation boundaries are shown for the AM coder and the AM/CA coder for 30 kbps for the excerpt Castanets. It can be seen that a higher coding efficiency is achieved as longer segments are chosen around the transients when AM coding templates are included. It was also found that when R-D optimal segmentation was used, there was still an advantage of using the onsets, i.e. improvements were still gained by allowing $n_l \neq 0$ in (2). Constraining $n_l = 0 \; \forall l$, i.e. reducing the model to that of [30, 33], led to shorter segments and a loss in perceived quality. The ability of the model to position onsets of the individual sinusoids at arbitrary positions within each segment has proven to be an important one. The effect of the rate-regularized estimation procedure is illustrated in Figure 8, where the rate-distortion curves of the AM coder for different regularization constants are shown for 2 s of claves. It can be seen that in the region 20-40 kbps, approximately 5 kbps can be saved compared to no regularization. Depending on the signal at hand, this result may vary.

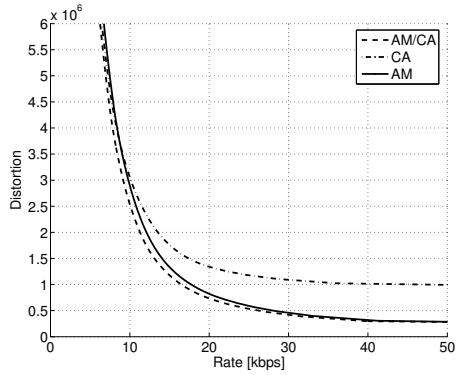

Figure 5: The rate-distortion curves of the CA coder (dash-dotted), the AM/CA coder (dashed) and the AM coder (solid) using a fixed segmentation for the glockenspiel.

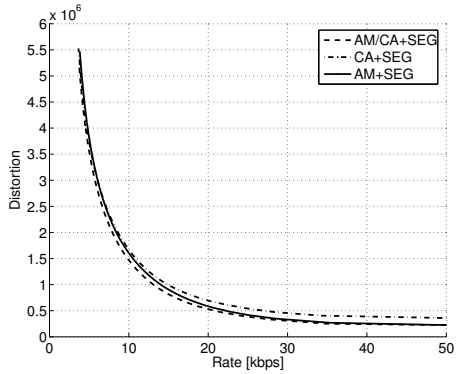

Figure 6: The rate-distortion curves of the CA+SEG coder (dash-dotted), the AM/CA+SEG coder (dashed) and the AM+SEG coder (solid) using R-D optimal segmentation for the glockenspiel.

7.2 Test Material

In order to evaluate the proposed method for parametric coding of transients, we conducted a formal listening test. In addition, we report our experience from informal listening tests to give the reader some indications as to the nature of the improvements that were made. In the informal and formal listening tests, the excerpts shown in Table 2 were used. These represent a wide variety of different types of signals, many of which are known to be critical excerpts in perceptual audio coding [37]. All the signals were monophonic and were 16 bit signals sampled at 48 kHz and they have a length of 6-12 s. Many more signals were used in the development, but these are the ones that have been tested extensively. In ITU-R BS.1534-1 [49] it is recommended to use excerpts that are known to be critical in testing of audio coding algorithms. Problematic transients by no means occur in all excerpts. Consequently, these tests are concerned mainly with excerpts that are known to be critical yet different of type. For example, the glockenspiel excerpt is very tonal and stationary for the most parts but has very steep attacks, while the castanet excerpt has very stochastic and strongly modulated characteristics. The excerpts 5 and 11 are pop music containing mixtures of multiple instruments and vocal.

7.3 Informal Listening Tests

Informal listening tests revealed that pre-echos are clearly reduced and that the transients are better modeled using the proposed model than with constant-amplitude sinusoids. For many signals, though, the improvements are fairly subtle since they are already handled well using constant-amplitude sinusoids. Often, the improvements are perceived as an increase of bandwidth of the coded signal. For critical excerpts, such

as castanets the improvements are clearly audible. The types of signals that benefit from the AM coder are signals that exhibit fast onsets, impulse-like signals, transitions between different stationary parts of signals, and percussive instruments. Any mixture of these types of signals with stationary ones may also benefit from it. It was also found that the AM coder improves the perceived quality of sinusoidally coded speech. Namely, the speech was found to suffer less from the tonal artifact often encountered in sinusoidal speech coding. Experiments showed that the AM coder proved R-D optimal for plosives, in transitions in pitch and in transitions between voiced and unvoiced sounds. For speech, it may also be beneficial to incorporate a model for frequency modulation [50]. Informal listening tests also revealed that the perceptual distortion measure (3) does not fully reflect the perceived improvement caused by the AM. For example, the relative improvement in terms of rate-distortion between the CA coder and the AM coder appears small for the castanets, while the perceived difference is large. This may be explained by the fact that the model [38] was derived for predicting the masking of sinusoidal component, and that the castanets are not very sinusoidal by nature unlike signals like the glockenspiel, claves and xylophone. The perceptual distortion measure (3) does, though, form a robust measure for estimation of model parameters and for the R-D optimization. When the R-D optimal segmentation is employed, the effects of the AM coder are less audible compared to the CA coder for excerpts where the signals exhibit fast onsets. Examples of this are glockenspiel and claves while for castanets, the combination of the AM coder and R-D optimal segmentation results in significant improvements. The use of variable bit-rate and R-D optimization has also been found to improve performance for transients for all the coders, since more bits can be allocated for critical signal parts, such as transients, this way.

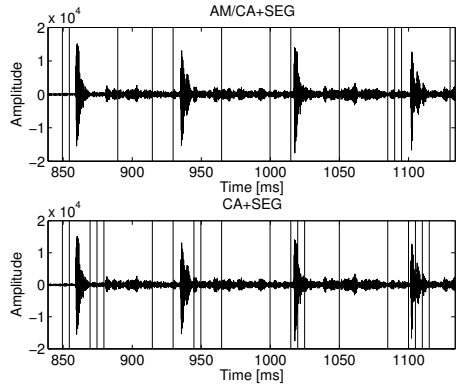

Figure 7: Example of R-D optimal segmentation boundaries (indicated by vertical lines) for castanets for the AM/CA+SEG coder (top) and the CA+SEG coder (bottom) operating at 30 kbps. Note that both the signals shown are the original.

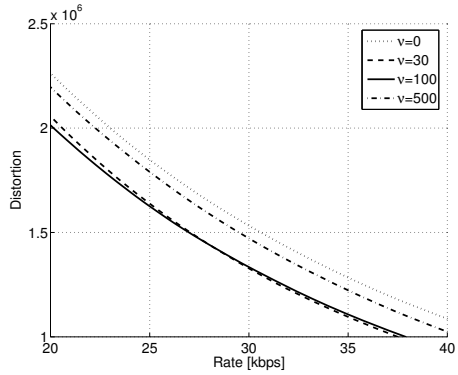

Figure 8: The rate-distortion curves of the AM coder for different regularization constants ν for claves optimized over 2 s.

7.4 MUSHRA Test

In order to quantify the improvements gained by the different methods for handling of transients, we use a subjective listening test. We use the MUSHRA test (MUlti-Stimulus test with Hidden Reference and Anchors) [49], which is a double blind test for subjective assessment of intermediate quality level of coding systems. For each excerpt, the listeners were asked to rank 8 differently processed versions relative to a known reference on a score from 0 to 100. These included the hidden reference (denoted HR), an anchor low-pass filtered at 7 kHz and an anchor low-pass filtered at 3.5 kHz (denoted Anchor 7 kHz and Anchor 3.5 kHz, respectively). The remaining 5 versions were the AM, CA, CA+SEG, AM+SEG and the AM/CA+SEG coders all operating at 30 kbps. In the MUSHRA test the hidden reference is used to verify the consistency of responses of subjects because a very high score is expected here. The anchors are included to be able to make comparisons between different listening tests and because they constitute a well-defined and simple signal modification. In order to limit the length of the listening test a representative subset of the excerpts listed in Table 2 was chosen. Nine expert listeners participated in the test (the authors not included). The test was performed on speakers in a listening room. As the proposed coders do no incorporate residual coding and are thus not complete parametric coders, a reference coder has not been included in this test. In MUSHRA tests the hidden references define known points on the scale. In Figure 9 the resulting MOS (Mean Opinion Score) scores of the different coder configurations averaged over all excerpts and listeners are shown. Since we are dealing with particular critical excerpts, it is of interest to investigate the performance for the individual excerpts. These are shown in Table 3 with the excerpt being identified by the number in Table 2. From Figure 9 we see that the AM/CA+SEG coder scores about 10 points higher at average than the CA+SEG coder, and more than 20 points higher

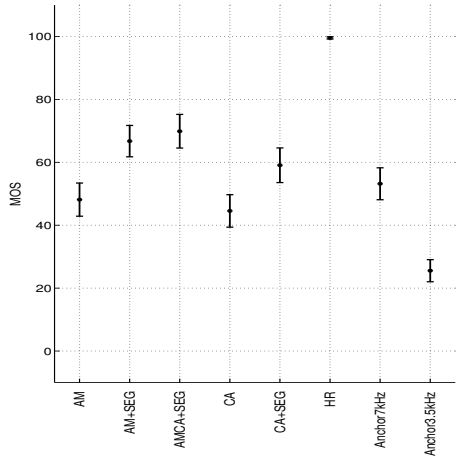

Figure 9: Results of the MUSHRA listening test. MOS scores for different coders averaged over all excerpts and all listeners. The error bars indicate the 95% confidence intervals.

than the CA coder. Although the AM coder does not seem to perform significantly better than the CA coder in this test, the AB preference test in [41] showed a significant preference for the AM/CA coder over the CA coder. In the table, it can be seen that for particular excerpts, such as the castanets (excerpt 9), there is a huge improvement in the combination of AM and the R-D optimal segmentation over the CA coder both with and without optimal segmentation, in fact the R-D optimal segmentation helps very little without the AM model. It can also be seen that there is a fairly small loss on average in the rate-regularized estimation procedure of the AM+SEG coder compared to the AM/CA+SEG, except for the glockenspiel (excerpt 3). Taking the confidence intervals into account, this difference is too small to be of any statistical significance. The reason for the fairly poor performance of the AM+SEG coder compared to the AM/CA+SEG coder for the glockenspiel is that the same regularization constant was used for processing all excerpts, and for the glockenspiel, this constant is not close to the optimal λ. It is interesting to note that the glockenspiel scores the highest among all excerpt. This is not surprising because the glockenspiel signal is very tonal and the AM model is well-suited for handling the non-stationary parts of this signal. This also holds for the very similar signals of SQAM, such as the claves, xylophone, triangle and others.

8 Discussion

As can be concluded from the listening test results, the proposed parametric coding of transients in combination with R-D optimal segmentation leads to a significant gain

Excerpt	1	3	5	7	9	10	11
AM	42	70	41	45	43	56	39
AM+SEG	67	79	58	71	66	68	58
AM/CA+SEG	65	92	62	68	72	71	59
CA	32	60	41	42	29	65	43
CA+SEG	47	84	64	63	35	65	55
HR	99	99	99	100	100	99	100
Anchor 7 kHz	47	66	56	62	47	42	52
Anchor 3.5 kHz	22	33	24	27	22	24	27

Table 3: Results of the MUSHRA listening test. MOS scores for different coder configurations for the individual excerpts.

in audio quality as compared to constant-amplitude sinusoidal coding. Switching between different window lengths and shapes or coders (e.g. [9, 18]) has traditionally been achieved by transient detection schemes. However, there may be a mismatch between the classification of transients and the R-D optimal coder. Based on R-D optimization and/or the rate-regularized estimation method robustness against such problems is gained, but this comes at the cost of additional complexity. We also note that the R-D optimal allocation scheme is similar to the so-called bit reservoir method for handling of transients (see [16]). Rate-distortion optimal allocation (variable rate) in itself does not, however, ensure that more bits are spent when transients are present. Rather, it spends the bits where most distortion can be reduced, and hence it depends on the appropriateness of the signal model.

The scores from the MUSHRA test reported here may be further improved by residual coding since noise components are not efficiently coded using sinusoids. Many parametric audio coders employ residual noise coding that only encodes a spectral and a coarse temporal envelope (e.g. [13, 51]). It is also possible to improve performance of parametric audio coders for transient signals by employing waveform approximating residual coding as done in [52, 53]. In such coding schemes, the residual coder may compensate for errors introduced by the sinusoidal coder.

Recently, preliminary results on linearization of the spectro-temporal psychoacoustical model [54] have been reported in [55]. Such a linearization results in a distortion measure that defines a norm and would thus be applicable to the AM estimation problem at the cost of increased complexity. Further, if such a measure is shown to reflect temporal aspects better than (3), this could lead to improved coding of transients as presented here as well as to more refined envelope dictionary design.

Compared to the singlebanded AM of e.g. [15], the model proposed in this paper has the advantage that different envelopes are allowed for different sinusoids, which is a particular advantage for mixtures of sources (see e.g. [34]). Some interesting parallels can be drawn to related work in audio coding. In [25] transient locations are modified

in order to achieve more efficient coding of transients. This is, in a sense, what is happening when the onsets are quantized, and seen in the light of [25], onsets should be estimated very precisely and then quantized jointly to a coarse grid. A successful tool in dealing with efficient coding of transients in transform coding is TNS [22]. TNS is based on linear predictive coding of transform coefficients. Since amplitude modulation may just as well be interpreted as a frequency domain filtering, there is a duality in TNS and AM. One conceptual difference between TNS and gain modification [23] as applied in transform coding on the one hand and AM as presented here on the other hand is that TNS and gain modification operate on the input and output signals and hence shape the noise, whereas in AM, the signal model is modified to fit the input signal.

9 Summary

In this paper, methods for efficient parametric coding of transient audio signals have been presented. We propose a specific model for handling of transients based on amplitude modulated sinusoids. In this model, each sinusoid is modulated by a different envelope known as a gamma envelope each being characterized by an onset, an attack and a decay parameter. These degrees of freedom have proven to be important in efficient coding of transients. Existing methods assume either that the modulating signal is the same for all components, that the onset always occurs at the start of a segment, or that no attack parameter is necessary. Combined with a constant-amplitude sinusoidal model, efficient coding of both stationary and transient signals is achieved using rate-distortion optimization based on a perceptual distortion measure. The rate-distortion optimization leads to optimal allocation and segmentation and therefore eliminates the need for transient detectors. Informal and formal listening tests reveal that for critical excerpts the combination of amplitude modulation and rate-distortion optimal segmentation leads to large improvements over a sinusoidal coder using only the optimal segmentation. This shows that segmentation techniques are not substitutes for good signal models.

Appendix A: Fourier Transform of Windowed Gamma Envelope

The estimation of model parameters and calculation of distortions require that the spectra of the windowed gamma envelopes are computed. Doing this by FFTs may be prohibitive for low complexity applications and storing them in memory may also not be feasible. Here, we instead derive a closed-form expression for generating the discrete-time Fourier transform directly in the frequency domain. The discrete-time Fourier transform of the windowed gamma envelope can be found from the following finite

sum:

$$Z_l(\omega) = \sum_{n=0}^{N-n_l-1} n^{\alpha_l} e^{-\beta_l n} w(n+n_l) e^{-j\omega(n+n_l)}, \tag{31}$$

with $w(n)$ being the tapered von Hann window (7). In finding the discrete Fourier transform we shall use the following transform pair:

$$n^a x(n) \leftrightarrow j^a \frac{\partial^a}{\partial \omega^a} X(\omega). \tag{32}$$

Assuming that $n_l < M-1$ and splitting the sum (31) up into three different sums having different window parts, we get

$$\begin{aligned} Z_l(\omega) = & \sum_{n=0}^{M-1-n_l} n^{\alpha_l} e^{-\beta_l n} v(n+n_l) e^{-j\omega(n+n_l)} \\ & + \sum_{n=M-n_l}^{N-M-1-n_l} n^{\alpha_l} e^{-\beta_l n} e^{-j\omega(n+n_l)} \\ & + \sum_{n=N-M-n_l}^{N-1-n_l} n^{\alpha_l} e^{-\beta_l n} v(n-N+2M+n_l) \\ & \times e^{-j\omega(n+n_l)}. \end{aligned} \tag{33}$$

with v(n) being the modified von Hann window in (8). Tedious calculations now lead to the following closed-form expression of the discrete-time Fourier transform of the windowed gamma envelopes:

$$\begin{aligned} Z_l(\omega) = j^{\alpha_l} \frac{\partial^{\alpha_l}}{\partial \omega^{\alpha_l}} \Bigg(& \frac{1}{2} e^{-j\omega n_l} \frac{1-(e^{-\beta_l - j\omega})^{M-n_l}}{1-e^{-\beta_l - j\omega}} \\ & - \frac{1}{4} e^{-j\omega n_l + j\frac{\pi}{M} n_l + j\frac{\pi}{2M}} \frac{1-(e^{-\beta_l - j\omega + j\frac{\pi}{M}})^{M-n_l}}{1-e^{-\beta_l - j\omega + j\frac{\pi}{M}}} \\ & - \frac{1}{4} e^{-j\omega n_l - j\frac{\pi}{M} n_l - j\frac{\pi}{2M}} \frac{1-(e^{-\beta_l - j\omega - j\frac{\pi}{M}})^{M-n_l}}{1-e^{-\beta_l - j\omega - j\frac{\pi}{M}}} \\ & + e^{-j\omega n_l} \frac{(e^{-\beta_l - j\omega})^{M-n_l} - (e^{-\beta_l - j\omega})^{N-M-n_l}}{1-e^{-\beta_l - j\omega}} \\ & + \frac{1}{2} e^{-j\omega n_l} \frac{(e^{-\beta_l - j\omega})^{N-M-n_l} - (e^{-\beta_l - j\omega})^{N-n_l}}{1-e^{-\beta_l - j\omega}} \\ & - \frac{1}{4} e^{-j\omega n_l + j\frac{\pi}{2M} - j\frac{\pi}{M}(N-n_l)} \\ & \times \frac{(e^{-\beta_l - j\omega + j\frac{\pi}{M}})^{N-n_l-M} - (e^{-\beta_l - j\omega + j\frac{\pi}{M}})^{N-n_l}}{1-e^{-\beta_l - j\omega + j\frac{\pi}{M}}} \end{aligned} \tag{34}$$

$$
\left. - \frac{1}{4} e^{-j\omega n_l - j\frac{\pi}{2M} + j\frac{\pi}{M}(N-n_l)} \right.
$$
$$
\left. \times \frac{(e^{-\beta_l - j\omega - j\frac{\pi}{M}})^{N-n_l-M} - (e^{-\beta_l - j\omega - j\frac{\pi}{M}})^{N-n_l}}{1 - e^{-\beta_l - j\omega - j\frac{\pi}{M}}} \right).
$$

In evaluating these expressions for particular parameter values and frequencies L'Hospital's rule must be used. For the coder presented in [41], where the window is simply a von Hann window with a fixed length, the corresponding expression is much simpler.

References

[1] P. Hedelin, "A tone oriented voice excited vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1981, pp. 205–208.

[2] L. Almeida and J. Tribolet, "Harmonic coding: A low bit-rate, good-quality speech coding technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 7, 1982, pp. 1664–1667.

[3] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(4), pp. 744–754, Aug. 1986.

[4] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4.

[5] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40(6), pp. 497–516, 1992.

[6] J. O. Smith and X. Serra, "Spectral Modelling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, 1990.

[7] K. N. Hamdy, M. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 1045–1048.

[8] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates," in *100th Conv. Aud. Eng. Soc.*, 1996, paper preprint 4179.

[9] S. N. Levine and J. O. Smith III, "A switched parametric & transform audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 985–988.

[10] T. S. Verma and T. H. Y. Meng, "A 6kbps to 85kbps scalable audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 877–880.

[11] H. Purnhagen and N. Meine, "HILN - The MPEG-4 Parametric Audio Coding Tools," in *IEEE International Symposium on Circuits and Systems*, 2000.

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

[13] A. C. den Brinker, E. G. P. Schuijers, and A. W. J. Oomen, "Parametric coding for high-quality audio," in *112th Conv. Aud. Eng. Soc.*, 2002, paper preprint 5554.

[14] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[15] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebart, "Advances in parametric coding for high-quality audio," in *114th Conv. Aud. Eng. Soc.*, 2003, paper preprint 5852.

[16] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88(4), pp. 451–515, Apr. 2000.

[17] L. L. Elliot, "Backward and forward masking of probe-tones of different frequencies," *J. Acoust. Soc. Am.*, vol. 34, pp. 1116–1117, 1962.

[18] B. Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," *Frequenz*, pp. 1033–1036, 1989.

[19] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech and Audio Processing*, pp. 646–655, 8(6) 2000.

[20] P. Prandoni, M. M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 2029–2032.

[21] M. Erne, G. Moschytz, and C. Faller, "Best wavelet-packet bases for audio coding using perceptual and rate-distortion criteria," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1999, pp. 909–912.

[22] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *101st Conv. Aud. Eng. Soc.*, 1996, paper preprint 4384.

[23] M. Link, "An attack processing of audio signals for optimizing the temporal characteristics of a low bit-rate audio coding system," in *95th Conv. Aud. Eng. Soc.*, 1993, paper preprint 3696.

[24] T. Vaupel, "Ein Beitrag zur transformationscodierung von Audiosignalen unter Verwendung der Methode der 'Time Domain Aliasing Cancellation (TDAC)' und einer Signalkompandierung im Zeitbereich," Ph.D. dissertation, Unversität-Gesamthochschule Duisburg, Germany, 1991.

[25] R. Vafin, R. Heusdens, and W. B. Kleijn, "Modifying transients for efficient coding of audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 3285–3288.

[26] S. N. Levine, T. S. Verma, and J. O. Smith III, "Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 1997, pp. 101–104.

[27] T. S. Verma and T. H. Y. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, 1998, pp. 3573–3576.

[28] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal models for audio modeling and coding," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Artificial Intelligence, V. Palade, R. J. Howlett, and L. C. Jain, Eds. Springer-Verlag, 2003, vol. 2773, pp. 1334–1342.

[29] J. Nieuwenhuijse, R. Heusdens, and E. F. Deprettere, "Robust Exponential Modeling of Audio Signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 3581–3584.

[30] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech and Audio Processing*, vol. 12(2), pp. 121–132, Mar. 2004.

[31] M. M. Goodwin, "Matching pursuit and atomic signal models based on recursive filter banks," *IEEE Trans. Signal Processing*, vol. 47(7), pp. 1890–1902, July 1999.

[32] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids," *IEEE Trans. Speech and Audio Processing*, vol. 12(2), pp. 110 – 120, Mar. 2004.

[33] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. van Huffel, "Perceptual audio modeling with exponentially damped sinusoids," *Signal Processing*, vol. 85, pp. 163–176, 2005.

[34] M. G. Christensen, S. van de Par, S. H. Jensen, and S. V. Andersen, "Multiband amplitude modulated sinusoidal audio modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2004, pp. 169–172.

[35] A. Aertsen and P. Johannesma, "Spectro-Temporal Receptive Fields of Audiotory Neurons in the Grass Frog. I. Characterization of tonal and natural stimuli," *Biol. Cybern.*, vol. 38, pp. 223–234, 1980.

[36] T. D. Rossing, *The Science of Sound*, 2nd ed. Addison-Wesley Publishing Company, 1990.

[37] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.

[38] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2001, pp. 1805 – 1808.

[39] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1445–1453, Sept. 1988.

[40] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[41] M. G. Christensen and S. van de Par, "Rate-distortion efficient amplitude modulated sinusoidal audio coding," in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2280–2284.

[42] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[43] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[44] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1999, pp. 981–984.

[45] M. M. Goodwin, "Adaptive Signal Models: Theory, Algorithms, and Audio Applications," Ph.D. dissertation, University of California, Berkeley, 1997.

[46] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. Academic Press, 1997.

[47] J. Jensen and R. Heusdens, "A comparison of differential schemes for low-rate sinusoidal audio coding," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2003, pp. 205–208.

[48] R. Heusdens, J. Jensen, W. B. Kleijn, V. kot, O. A. Niamut, S. van de Par, N. H. van Schijndel, and R. Vafin, "Bit-rate scalable intra-frame sinusoidal audio coding based on rate-distortion optimisation," *J. Audio Eng. Soc.*, 2005, submitted.

[49] *ITU-R BS.1534*, ITU Method for subjective assessment of intermediate quality level of coding system, 2001.

[50] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Processing*, vol. 41(10), pp. 3024–3051, Oct. 1993.

[51] M. M. Goodwin, "Residual modeling in music analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1996, pp. 1005–1008.

[52] F. Riera-Palou, A. C. den Brinker, and A. J. Gerrits, "A hybrid parametric-waveform approach to bistream scalable audio coding," in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2250–2254.

[53] R. Vafin and W. B. Kleijn, "Towards optimal quantization in multistage audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2004, pp. 205–208.

[54] 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. Am.*, vol. 99(6), pp. 3615–3622, June 1996.

[55] J. Plasberg, D. Zhao, and W. B. Kleijn, "Sensitivity matrix for a spectro-temporal auditory model," in *Proc. XII European Signal Processing Conf. (EUSIPCO)*, 2004, pp. 1673–1676.

Paper D

Computationally Efficient Amplitude Modulated Sinusoidal Audio Coding using Frequency-Domain Linear Prediction

Mads Græsbøll Christensen and Søren Holdt Jensen

The paper has been submitted to
IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006.

Abstract

A method for amplitude modulated sinusoidal audio coding is presented that has low complexity and low delay. This is based on a subband processing system, where, in each subband, the signal is modeled as an amplitude modulated sum of sinusoids. The envelopes are estimated using frequency-domain linear prediction and the prediction coefficients are quantized. As proof of concept, we evaluate different configurations in a subjective listening test, and this shows that the proposed method offers significant improvements in sinusoidal coding. Furthermore, the properties of the frequency-domain linear prediction-based envelope estimator are analyzed.

1 Introduction

Parametric coding of audio and speech has received considerable attention in the research community and standardization bodies in recent years [1–3]. In order to achieve good performance at low bit-rates, parametric coding relies on signal models that describe the signal in few physically meaningful parameters. Parametric audio coders perform extremely well when the signal fits the signal model. However, when this is not the case, the coded signal may be of very low perceived quality. This can be observed from subjective listening tests where the scores may vary greatly depending on the signal (see e.g. [4]). In [4] it was shown that even when using rate-distortion optimal segmentation [5], constant-amplitude sinusoids do not lead to satisfactory results for critical transients excerpts such as those from SQAM [6]. It was demonstrated that an amplitude modulated (AM) sinusoidal audio coder lead to significant improvements over a sinusoidal coder for such signals. The coder was based on an analysis-by-synthesis parameter estimation procedure using a perceptual distortion measure. As a consequence, the coder suffered from high complexity and delay. Further, it was also shown in [4] that significant improvements are gained by the combination of rate-distortion optimal segmentation and amplitude modulated sinusoidal audio coding, i.e. that model adaptation and flexible segmentation are complementary tools. The rate-distortion optimal segmentation requires that all possible segment lengths at different starting positions be coded. This, of course, adds considerable complexity and delay, which may be prohibitive for some applications. For example, the MPEG-4 Low Delay Audio Coder [7] does not use block switching and minimizes the use of the bit reservoir due to the delay associated with these methods. A powerful and successful method for efficient coding of transients in the context of transform coding is the so-called temporal noise shaping (TNS) [8], which is part of the MPEG-2/4 AAC and is used in the Low Delay Audio Coder instead of block switching [7]. In TNS, a coding gain is achieved for transient signals by predictive coding of transform coefficients, and the optimal linear predictor can be derived efficiently in the frequency domain using standard methods [9]. While linear prediction has found use in predictive coding of speech, it is also a widely used

method for spectral estimation of auto-regressive (AR) stochastic processes. Likewise, TNS can be interpreted as either a method for predictive coding in the frequency domain or as an envelope estimator, or, in modulation theoretical terms, a demodulator. Aside from being an envelope estimator, the frequency-domain linear predictor is also an efficient parametric representation of the envelope that can be quantized using well-known methods.

In this paper, we explore an alternative amplitude modulated sinusoidal coding technique based on frequency-domain linear prediction (FDLP) that has considerably lower complexity and delay than [4]. Further, we also analyze the properties of the FDLP-based envelope estimator. We apply the sinusoidal coding in the subbands of a critically sampled filterbank. The advantage of doing this is twofold. Firstly, it has a lower computational complexity than the full-band system and, secondly, it allows for different envelopes in the individual subbands, which may be desirable for some, but by no means all, signals [4, 10]. The sinusoidal parameters are found, given the subband envelopes, using matching pursuit based on a perceptual distortion measure.

The remaining part of this paper is organized as follows: Section 2 contains an overview of the proposed system. The envelope estimator and its properties are presented in Section 3 and subsequently the matching pursuit algorithm used for sinusoidal parameter estimation is treated in Section 4. In Sections 5 and 6 implementation details and experimental results are presented, and, finally, Section 7 concludes on the work.

2 System Overview

The method proposed in this paper is implemented in a system that consists of an analysis and a synthesis system. In the analysis system the input signal is first split into Q critically sampled subbands using a uniform analysis filterbank. Then in each subband, an envelope is estimated using FDLP and the associated parameters are quantized. Given the subband envelopes a number of sinusoidal parameters are extracted and quantized and finally all parameters are then entropy coded. In the synthesis system, the parameters are reconstructed and the subband signals are synthesized using overlap-add. Finally, the signal is reconstructed using a synthesis filterbank that combined with the analysis filterbank has perfect reconstruction. Now, let us introduce some definitions and the notation. First, we define the subband signal for subband q as $x_q(n) = \sum_{m=0}^{M-1} h_q(m)x(n-m)$ for $n = 0, \ldots, N-1$ with $h_q(n)$ being the impulse response of the qth analysis filter of length M. From these subband signals, the input signal can be reconstructed (with a delay d) as $\sum_{q=1}^{Q} \sum_{m=0}^{M-1} g_q(m)x_q(n-m) = x(n-d)$ using the impulse responses of the synthesis filters $g_q(n)$. In the envelope estimation and in the sinusoidal parameter estimation, we will make use of the so-called discrete-time analytic signal for a particular segment. For $n = 0, \ldots, K-1$ with $K = N/Q$ this is defined as

$$a_q(n) = x_q(Qn) + j\mathcal{H}\left\{x_q(Qn)\right\}, \tag{1}$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. Here we have assumed that the segment length N is an integer multiple of Q. Note that the calculation of the analytic signal may be integrated into the filterbank implementation since $h_q(n) * (x(n) + j\mathcal{H}\{x(n)\}) = x(n) * (h_q(n) + j\mathcal{H}\{h_q(n)\})$. The model of the analytic subband signal used in this paper can be written as the following sum of sinusoids where the sinusoids are modulated by the (complex) amplitude modulating signal $\hat{\gamma}_q(n)$,

$$\hat{a}_q(n) = \sum_{l=1}^{L_q} \hat{\gamma}_q(n) A_{q,l} e^{j\omega_{q,l}n + j\phi_{q,l}}, \tag{2}$$

where each sinusoid is characterized by a frequency $\omega_{q,l}$, a phase $\phi_{q,l}$ and an amplitude $A_{q,l}$. This signal model is built using a matching pursuit algorithm [11] based on the perceptual distortion measure presented in [12] and a redundant dictionary consisting of modulated complex sinusoidal components. This dictionary can be seen as being signal adaptive since the amplitude modulating signal $\hat{\gamma}_q(n)$ varies with the signal over time and subbands.

In the decoder, the subband signal is recovered by taking the real-value of (2), upsampling the signal by a factor of Q and subsequent filtering by the synthesis filter $g_q(n)$.

3 Envelope Estimation

In this section we briefly present the main results of the envelope estimator based on the FDLP principle first introduced in [8] as temporal noise shaping for transform coding. Further, we also provide some additional analysis of the properties of this estimator. It must be emphasized that the application of FDLP considered here is fundamentally different from that of [8]. First, we define the Fourier transform of the analytic signal for $k = 0, \ldots, K-1$ as

$$A_q(k) = \sum_{n=0}^{K} a_q(n) e^{-j2\pi \frac{k}{K} n}. \tag{3}$$

We then write the frequency-domain prediction error $E_q(k)$ as a linear combination of $A_q(k)$ having the I (complex) prediction coefficients $b_i \in \mathbb{C}$:

$$E_q(k) = A_q(k) - \sum_{i=1}^{I} b_i A_q(k-i). \tag{4}$$

The optimal prediction coefficients are then found in such a way that the squared prediction error is minimized, i.e.,

$$\{b_i\} = \arg\min \sum_{k=0}^{K-1} |A_q(k) - \sum_{i=1}^{I} b_i A_q(k-i)|^2, \tag{5}$$

which can be solved efficiently using well-known methods [9]. Then, by taking the Fourier transform of the squared instantaneous envelope, we get the spectral autocorrelation sequence estimate $C_q(\tau)$:

$$\sum_{n=0}^{K-1} |a_q(n)|^2 e^{-j2\pi\frac{\tau}{K}n} = \frac{1}{K}\sum_{k=0}^{K-1} A_q(k)A_q^*(k-\tau) \tag{6}$$

$$= C_q(\tau), \tag{7}$$

from which the prediction coefficients also can be found. Taking the inverse Fourier transform of both sides of (4), we get

$$e_q(n) = a_q(n)\left[1 - \sum_{i=1}^{I} b_i e^{j2\pi\frac{i}{K}n}\right]. \tag{8}$$

Rearranging this, we get the (complex) envelope estimate for $n = 0, \ldots, K$,

$$\hat{\gamma}_q(n) = \frac{a_q(n)}{e_q(n)} = \frac{1}{1 - \sum_{i=1}^{I} b_i e^{j2\pi\frac{i}{K}n}}. \tag{9}$$

Specifically, the squared instantaneous envelope estimate is

$$|\hat{\gamma}_q(n)|^2 = \frac{1}{|1 - \sum_{i=1}^{I} b_i e^{j2\pi\frac{i}{K}n}|^2}. \tag{10}$$

As the prediction filter is minimum-phase, the phase of $\hat{\gamma}_q(n)$ can be determined uniquely from $\log|\hat{\gamma}_q(n)|$ since they form a Hilbert transform pair, i.e.,

$$\angle\hat{\gamma}_q(n) = \mathcal{H}\{\log|\hat{\gamma}_q(n)|\}. \tag{11}$$

Using Parseval's theorem, we can write the minimization in (5) of the sum of the squared prediction in the time domain:

$$\min \sum_{k=0}^{K-1} |E_q(k)|^2 = \min \sum_{n=0}^{K-1} \frac{|a_q(n)|^2}{|\hat{\gamma}_q(n)|^2}. \tag{12}$$

From these equations, we can make a number of interesting observations. From (12) we see that minimizing the prediction error a least-squares fashion corresponds to minimizing the sum of the ratio between the squared instantaneous envelope and the estimate. The frequency-domain linear predictor models the time-domain envelope in exactly the same way as the time-domain linear predictor models the spectral envelope, and, hence, they share the same properties and suffer from the same problems (e.g. the cancellation of errors and overemphasis on peaks [9]). One notable property is that the envelope estimate converges to the squared instantaneous envelope as the model order I grows.

From (7) and (12) we see that a decorrelation of the transform coefficients results in a flattening of the squared instantaneous envelope since $C_q(\tau) = 0$ for $\tau \neq 0$ implies a flat envelope. It can easily be shown that the squared instantaneous envelope of two sinusoids that are modulated by the same signal contains cross-terms that are due to the sinusoidal carriers (see [10]). In sinusoidal modeling these cross-terms are modeled by the sinusoids and should hence not be captured by the envelope estimator. For sinusoids that are well-separated in frequency these cross-terms will occur as long-term correlation in $C_q(\tau)$, and hence FDLP has the (at least in this particular application) undesirable property that it will seek to model these cross-terms. Consequently, the model order should be chosen sufficiently low such that this does not happen and this order cannot, contrary to common practice in transform coding, simply be chosen from the prediction gain. Moreover, the envelope estimator will fail when the sinusoids and the modulating signal are not well-separated in frequency (since Bedrosians theorem [13] does not hold in this case) or when sinusoids are closely spaced [10].

4 Subband Matching Pursuit

The individual sinusoidal parameters, i.e. frequencies, phases and amplitudes, are found in each subband using a psychoacoustic matching pursuit [11] given the subband envelopes $\hat{\gamma}_q(n)$. The subband envelope adapts the dictionary to the subband signal, and, as a consequence, a higher rate of convergence, in terms of the distortion as a function of the number of components, can be achieved. In each iteration, with i being the iteration index, the algorithm operates on the F point Fourier transform of the subband residuals $R_{q,i}(k)$ which are initialized for $k = 0, \ldots, F-1$ as

$$R_{q,1}(k) = \sum_{n=0}^{K-1} w(n) a_q(n) e^{-j2\pi \frac{k}{F} n}, \tag{13}$$

with $w(n)$ being the analysis/synthesis window. The algorithm finds in each iteration the subband and the parameters that minimize the weighted squared absolute value of the Fourier transform of the residual, i.e., $D_{q,i} = \sum_{k=0}^{F-1} P_q(k) |R_{q,i}(k)|^2$, where $P_q(k)$ is a perceptual weighting function for the frequency region associated with subband q. This weighting function is derived, for each segment, from the auditory masking model presented in [12]. The combination of frequency (index) $\hat{f}$ and subband $\hat{q}$ that minimizes the perceptual distortion are chosen as:

$$\{\hat{q}, \hat{f}\} = \arg\max_{\{q,m\}} \frac{|\Psi_q(m)|^2}{\Phi_q(m)}, \tag{14}$$

with the numerator containing the inner product

$$\Psi_q(m) = \sum_{k=0}^{F-1} P_q(k) Z_q^*(k-m) R_{q,i}(k), \tag{15}$$

and the denominator the norm

$$\Phi_q(m) = \sum_{k=0}^{F-1} P_q(k) Z_q^*(k-m) Z_q(k-m). \tag{16}$$

$Z_q(k)$ is defined as the Fourier transform of the windowed subband envelope, i.e.,

$$Z_q(k) = \sum_{n=0}^{K-1} w(n) \hat{\gamma}_q(n) e^{-j2\pi \frac{k}{F} n}. \tag{17}$$

The optimum phase and amplitude associated with the estimated complex sinusoid of frequency $\hat{f}$ in subband $\hat{q}$ can be found as

$$A_{\hat{q},i} e^{j\phi_{\hat{q},i}} = \frac{\Psi_{\hat{q}}(\hat{f})}{\Phi_{\hat{q}}(\hat{f})}. \tag{18}$$

Finally, having found the subband, frequency, phase and amplitude we update the Fourier transform of that subband residual as

$$R_{\hat{q},i+1}(k) = R_{\hat{q},i}(k) - A_{\hat{q},i} e^{j\phi_{\hat{q},i}} Z_{\hat{q}}(k - \hat{f}). \tag{19}$$

Note how the numerators of equations (18) and (14) contain the same inner product. It can be seen from the following that, like for the constant-amplitude case treated in [11], these inner products can be efficiently computed for different m using FFTs:

$$\Psi_q(m) = \sum_{n=0}^{K-1} v_q(n) w(n) \hat{\gamma}_q^*(n) e^{-j2\pi \frac{m}{F} n}, \tag{20}$$

with $v_q(n) = \sum_{k=0}^{F-1} P_q(k) R_{q,i}(k) e^{j2\pi \frac{k}{F} n}$. Similarly, the denominator of equations (18) and (14) can be found using Fourier transforms:

$$\Phi_q(m) = \sum_{n=0}^{K-1} \left[\sum_{k=0}^{F-1} |Z_q(k)|^2 e^{-j2\pi \frac{k}{F} n} \right] p_q(n) e^{-j2\pi \frac{m}{F} n}, \tag{21}$$

with $p_q(n)$ being the inverse Fourier transform of the perceptual weighting function, i.e., $p_q(n) = \sum_{k=0}^{F-1} P_q(k) e^{j2\pi \frac{k}{F} n}$. Finally, we note that since the subband signals are orthogonal, the total distortion is simply the sum of the subband distortions and can hence be subject to rate-distortion optimization. It follows from the perfect reconstruction of the filterbank and the convergence of the matching pursuits in the subbands that the entire system will converge.

Statistic	AM	CA	Anchor 1	Anchor 2	HR
Mean	62	54	29	54	95
Conf. ($\pm$)	7.4	6.6	4.0	5.8	2.8

Table 1: Results of MUSHRA test. Scores for all excerpts and listeners (means and 95% confidence intervals).

5 Implementation Details

In assessing the improvement that the higher update-rate of the amplitude, i.e. amplitude modulation, results in, we compare two different configurations of the proposed system: The first uses only constant-amplitude sinusoids (denoted CA) while the second uses multiband amplitude modulation with Q=8 (denoted AM). The optimal segment length has been determined empirically by informal listening tests to be about 35 ms for the CA configuration using a von Hann window with 50% overlap. This segment length was also used for the AM configuration. The frequencies and amplitudes are quantized using the logarithmic quantizer described in [4] while the phases are quantized uniformly using 5 bits. At average this results in approximately 15 bits per sinusoidal component. The complex prediction coefficients were quantized by mapping the reflection coefficients to the log-area ratios which were then quantized uniformly. The associated rate has been estimated from the entropy of the quantization indices, which resulted in approximately 9 bits per complex prediction coefficient, and a 5th order complex prediction filter was used in the simulations. Both configurations were set to run at 30 kbps. In order to achieve efficient coding of stationary sinusoids, the envelope is not used when it has a correlation coefficient of more than 90% with a constant envelope. Alternatively, the rate-distortion optimization-based coder switching architecture proposed in [4] can be used for this. Subband FFTs of 1024 points were used for the AM configuration while the CA configuration uses 8192 point FFTs in the matching pursuit. Further, we have used an FFT-based implementation of the Hilbert transform. In practice we have found the orthogonality between subbands not to be of critical importance for the task at hand. Hence, we have used a uniform 8-band pseudo QMF filterbank for the AM configuration with a prototype filter of length 512.

6 Results and Discussion

In Figure 1 we illustrate the order selection problem of the envelope estimator for two sinusoids $2\cos(2\pi 0.1n) + \sin(2\pi 0.11n + \pi/3)$. It can be seen that the FDLP models the cross-terms that are due to the carriers and that the low-order model is better than the high-order model when cross-terms are present.

The subband processing has been verified to produce good results compared to a

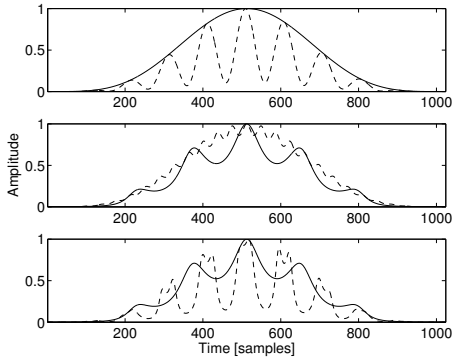

Figure 1: Illustration of the problem of cross-terms in the squared instantaneous envelope for multiple sinusoids. The top panel shows the true squared envelope (solid) and the squared instantaneous envelope (dashed). The squared instantaneous envelope of one sinusoid estimated using a 5th order predictor (solid), and using a 25th order predictor (dashed) are shown in the middle panel. And similarly, in the bottom panel, for two sinusoids estimated using a 5th order predictor (solid), and using a 25th order predictor (dashed).

full-band system using constant-amplitude sinusoids, i.e. no AM. This verifies that the subband processing does not introduce any noticeable artifacts, although it has some inherent drawbacks. There is a significant reduction of complexity of the subband system compared to the full-band system. Where the full-band system would require FFTs of size F, the subband system requires FFTs of size F/Q and only one subband has to be updated per iteration of the matching pursuit.

As proof of concept of the proposed method, we use a subjective listening test. Specifically, we use the MUSHRA test [14] for quantifying the improvements of the AM configuration compared to the CA configuration. For each excerpt, the listeners were asked to rank 5 differently processed versions relative to a known reference on a score from 0 to 100. These included the hidden reference (denoted HR), an anchor low-pass filtered at 3.5 kHz and an anchor low-pass filtered at 7 kHz (denoted Anchors 1 and 2). The remaining two versions were the different configurations, AM and CA. The excerpts were 5 different critical 10 s transient (mono) signals from SQAM [6], namely castanets, claves, glockenspiel, triangle and xylophone. 12 inexperienced listeners participated and the test was conducted using headphones. In Table 1 the results of the listening test are shown. As can be seen from the anchors, the sometimes large confidence intervals can largely be attributed to variations over the excerpts and listeners. Testing instead for the differences in scores, the mean of the difference between the AM and the CA configuration was found to be 8.3 with a 95% confidence interval of ± 7.4. Since the confidence interval does not include zero, we conclude that the AM configuration performs significantly better than the CA configuration. In interpreting the results it should noted that due to the temporal integration in the human auditory system, events of a short duration, such as onsets, may only result in small improve-

ments in scores. We also note that, as shown in [4], both the CA and AM configurations may be further improved and even combined using rate-distortion optimal segmentation and allocation [5] at the cost of significantly increased delay and complexity.

7 Conclusion

We have presented a method for amplitude modulated sinusoidal audio coding based on frequency-domain linear prediction for estimation and efficient coding of time-domain envelopes. This has been found, in a subjective listening test, to improve on sinusoidal coding for critical transient signals. Further, the properties of the envelope estimator have been analyzed and it has been demonstrated that special care must be taken in selecting the model order.

References

[1] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates," in *100th Conv. Aud. Eng. Soc.*, 1996, paper preprint 4179.

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

[3] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebart, "Advances in parametric coding for high-quality audio," in *114th Conv. Aud. Eng. Soc.*, 2003, paper preprint 5852.

[4] M. G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(4), July 2006.

[5] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech and Audio Processing*, pp. 646–655, 8(6) 2000.

[6] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.

[7] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "MPEG-4 Low Delay Audio Coding based on the AAC Codec," in *106th Conv. Aud. Eng. Soc.*, May 1999, paper preprint 4929.

[8] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *101st Conv. Aud. Eng. Soc.*, 1996, paper preprint 4384.

[9] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, vol. 63(4), pp. 561–580, Apr. 1975.

[10] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal models for audio modeling and coding," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Artificial Intelligence, V. Palade, R. J. Howlett, and L. C. Jain, Eds. Springer-Verlag, 2003, vol. 2773, pp. 1334–1342.

[11] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[12] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2004.

[13] E. Bedrosian, "A product theorem for Hilbert transforms," in *Proc. IEEE*, vol. 51(1), May 1963, pp. 868–869.

[14] *ITU-R BS.1534*, ITU Method for subjective assessment of intermediate quality level of coding system, 2001.

Paper E

Linear AM Decomposition for Sinusoidal Audio Coding

Mads Græsbøll Christensen, Andreas Jakobsson, Søren Vang Andersen, and Søren Holdt Jensen

The paper has been published in
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 165–168, 2005.

Abstract

In this paper, we present a novel decomposition for sinusoidal audio coding using amplitude modulation of sinusoids via a linear combination of arbitrary basis vectors. The proposed method, which incorporates a perceptual distortion measure, is based on a relaxation of a non-linear least squares minimization. It offers benefits in the modeling of transients in audio signals. We compare the decomposition to constant-amplitude sinusoidal coding using rate-distortion curves and listening tests. Both indicate that, at the same bit-rate, perceptually significant improvements can be achieved using the proposed decomposition.

1 Introduction

The problem of decomposing a signal into amplitude modulated sinusoids is encountered in many different applications, for example in parametric audio coding (see, e.g., [1]) where modulated sinusoidal models are of interest for handling transients. Even when dynamic time segmentation [2, 3] is employed, there is a need for efficient modeling of transients. In [4], it was shown that perceptually significant improvements can be achieved by applying amplitude modulation (AM) in a frequency dependent way as opposed to single-banded AM (see, e.g., [5]). Furthermore, it was shown in [6] that frequency dependent AM achieves lower distortions compared to constant-amplitude (CA) sinusoidal coding at the same rate. Sinusoidal modeling using both amplitude and frequency modulation, in the form of a linear combination of basis vectors such as low-order polynomials, has been explored for a variety of applications (see, e.g., [7, 8]). Although such models perform well for slowly evolving signals like voiced speech, they do not handle the transients often encountered in audio signals well.

In this paper, we extend the work in [4, 6] by introducing a signal decomposition based on a set of preselected, linearly independent, real-valued basis vectors that describe the amplitude modulating signal. Furthermore, we examine how to incorporate such a decomposition in parametric audio coding, especially noting that it is not always efficient in terms of rate and distortion to use the AM technique. The rest of the paper is organized as follows: In Section 2, both the signal decomposition and the solution to the associated minimization problem are presented, followed in Section 3 with the incorporation of a perceptual distortion measure. Section 4 describes sinusoidal audio coding using the proposed AM decomposition. Experimental results are presented in Section 5, and Section 6 concludes on our work.

2 Proposed Decomposition

In the proposed decomposition, the signal of interest is modeled as a sum of amplitude modulated sinusoids, i.e.,

$$x(n) = \sum_{l=1}^{L} \gamma_l(n) \cos(\omega_l n + \phi_l), \tag{1}$$

where ω_l and ϕ_l denote the lth carrier frequency and phase, respectively, and $\gamma_l(n)$ is the amplitude modulating signal formed as the linear combination

$$\gamma_l(n) = \sum_{i=1}^{I} b(n,i) c_{i,l}, \tag{2}$$

where $b(n,i)$ and $c_{i,l}$ denote the ith basis function evaluated at time instance n and the (i,l)th AM coefficient, respectively. We will here assume that the L carrier frequencies are distinct, so that $\omega_k \neq \omega_l$ for $k \neq l$. The additional flexibility in (1), as compared to the traditional constant-amplitude models with $\gamma_l(n) = A_l$, gives improved modeling of transient segments. We note that the constant-amplitude model is a special case of the modulated model, with the amplitude modulating signal being DC. Let $x_a(n)$ denote the discrete-time "analytical" signal constructed from $x(n)$ by removing the negative frequency components, such that the resulting signal may be down-sampled by a factor two without loss of information [9] provided that there is little or no signal of interest near 0 and π. The signal model $x_a(n)$ can then be written as

$$x_a(n) = \sum_{l=1}^{L} \sum_{i=1}^{I} b(n,i) c_{i,l} e^{j\omega_l n + j\phi_l} \tag{3}$$

Choosing N to be even, and introducing

$$\mathbf{x}_a = \begin{bmatrix} x_a(1) & x_a(3) & \cdots & x_a(N-1) \end{bmatrix}^T, \tag{4}$$

where $(\cdot)^T$ is the transpose operator, the down-sampled discrete-time "analytical" signal may be put into matrix-vector notation

$$\mathbf{x}_a = [(\mathbf{BC}) \odot \mathbf{Z}]\, \mathbf{a}, \tag{5}$$

where $\odot$ denotes the Schur-Hadamard product, i.e., $[\mathbf{E} \odot \mathbf{F}]_{kl} = [\mathbf{E}]_{kl}[\mathbf{F}]_{kl}$, with $[\mathbf{E}]_{kl}$ being the (k,l)th element of $\mathbf{E}$. Further, $\mathbf{Z} \in \mathbb{C}^{N/2 \times L}$ with $L < N/2$ is constructed from the L complex carriers, i.e., $[\mathbf{Z}]_{kl} = e^{j\omega_l(2k-1)}$, $\mathbf{a} = \begin{bmatrix} e^{j\phi_1} & \cdots & e^{j\phi_L} \end{bmatrix}^T$. The amplitude modulating signal is written using the known AM basis vectors, $[\mathbf{B}]_{kl} = b(2k-1,l)$, and the corresponding coefficients, $[\mathbf{C}]_{kl} = c_{k,l}$. Here, $\mathbf{B} \in \mathbb{R}^{N/2 \times I}$ with

$I < N/2$ and $\mathbf{C} \in \mathbb{R}^{I \times L}$. The problem of interest is given a measured signal, $y(n)$, find $x(n)$ such that

$$\min_{\mathbf{C},\{\phi_k\},\{\omega_k\}} \sum_{n=1}^{N} |y(n) - x(n)|^2 \tag{6}$$

or, equivalently,

$$\min_{\mathbf{C},\{\phi_k\},\{\omega_k\}} \|\mathbf{y}_a - \mathbf{x}_a\|_2^2 \tag{7}$$

where $\mathbf{y}_a$ is formed similar to $\mathbf{x}_a$, and $\|\cdot\|_2$ denotes the 2-norm. This problem is nonlinear in the frequencies $\{\omega_k\}_{k=1}^L$, and is thus called a nonlinear least squares (NLS) minimization. Typically, this type of problem requires a multidimensional minimization which is computationally infeasible in most situations. For the sinusoidal estimation problem, several suboptimal approaches based on relaxation of the original problem have been suggested to reduce the computational complexity of the minimization, such as the greedy matching pursuit [10] or recursive methods such as RELAX [11]. Herein, we propose an iterative method for the minimization of (7), reminiscent to both the above mentioned methods. The suggested method exploits the fact that for given $\{\omega_k\}_{k=1}^L$, the minimization problem with respect to $\mathbf{C}$ for fixed $\{\phi_k\}_{k=1}^L$ is quadratic, and conversely the minimization of $\{\phi_k\}_{k=1}^L$ for fixed $\mathbf{C}$. We propose to iteratively find $\mathbf{C}$ and $\{\phi_k\}_{k=1}^L$, minimizing the residual for each frequency in a given finite set of frequencies, Ω. Let

$$\mathbf{c}_k = \begin{bmatrix} c_{1,k} & \cdots & c_{I,k} \end{bmatrix}^T. \tag{8}$$

At iteration k, assuming the $k-1$ carriers and corresponding coefficients known (i.e., found in prior iterations), we find, for each frequency $\omega \in \Omega$, the model parameters ϕ_k and $\mathbf{c}_k$, minimizing the residual for that particular frequency. The kth carrier is then found as the parameter set minimizing the residual over Ω, i.e.,

$$\hat{\omega}_k = \arg\min_{\omega \in \Omega} \|\mathbf{r}_k - \mathbf{D}_k e^{j\phi_k} \mathbf{B}\mathbf{c}_k\|_2^2, \tag{9}$$

where $\mathbf{D}_k$ is the diagonal matrix constructed from the kth carrier, with $z_k = e^{j\omega_k}$, i.e.,

$$\mathbf{D}_k = \operatorname{diag}\left(\begin{bmatrix} z_k^1 & z_k^3 & \cdots & z_k^{N-1} \end{bmatrix}\right). \tag{10}$$

Further,

$$\mathbf{r}_k = \begin{bmatrix} r_k(1) & r_k(3) & \cdots & r_k(N-1) \end{bmatrix}^T \tag{11}$$

contains the kth residual, obtained as

$$r_k(n) = y_a(n) - \sum_{l=1}^{k-1} \sum_{i=1}^{I} b(n,i) \hat{c}_{i,l} e^{j\hat{\omega}_l n + j\hat{\phi}_l}. \tag{12}$$

For each frequency ω, we iteratively solve for ϕ_k and $\mathbf{c}_k$ (with superscript (p) denoting the pth iteration of the alternating minimization); for given $\hat{\mathbf{c}}_k^{(p-1)}$,

$$\hat{\phi}_k^{(p)} = \angle \left\{ \sum_{\substack{n=1,\\ n \text{ odd}}}^{N} \sum_{i=1}^{I} b(n,i) \hat{c}_{i,l}^{(p-1)} e^{-j\omega n} r_k(n) \right\}. \tag{13}$$

Given $\hat{\phi}_k^{(p)}$, the minimization wrt. the AM coefficients reduces to

$$\hat{\mathbf{c}}_k^{(p)} = \mathbf{B}^{\dagger} \mathbf{u}_k^{(p)}, \tag{14}$$

with

$$\mathbf{B}^{\dagger} = \left(\mathbf{B}^T \mathbf{B}\right)^{-1} \mathbf{B}^T, \tag{15}$$

which can be pre-computed. The vector $\mathbf{u}_k^{(p)}$ is defined as

$$\mathbf{u}_k^{(p)} = \left[\begin{array}{cccc} u_k^{(p)}(1) & u_k^{(p)}(3) & \cdots & u_k^{(p)}(N-1) \end{array} \right]^T, \tag{16}$$

which is the real part (recall that $c_{i,l} \in \mathbb{R}$) of the residual shifted towards DC by the carrier, i.e.,

$$u_k^{(p)}(n) = \text{Re} \left\{ r_k(n) e^{-j\omega n - j\hat{\phi}_k^{(p)}} \right\}. \tag{17}$$

The parameters in (13) and (14) are then found alternately, given the other, until some stopping criterion is reached. For a given ω the problem is convex, and the algorithm converges to a global maximum. Hence, the 2-norm of the residual is a non-increasing, convex function of the number of iterations. We note that for the special case of constant amplitude (DC basis), the estimates (9), (13) and (14) reduce to those of a matching pursuit [10] with complex sinusoids.

3 Incorporating Perceptual Distortion

It is well-known that the 2-norm error measure does not correlate well with human sound perception. The problem of finding a suitable distortion measure is one of computational complexity and mathematical convenience and tractability. On one hand, we would like to have a measure that takes as much as possible of the processing in the human auditory system into account, while on the other hand, we would like to have a measure that defines a mathematical norm and leads to efficient, simple estimators and quantizers. Here we apply the perceptual distortion measure presented in [12]. For a particular segment, the distortion D can be written as

$$D = \int_{-\pi}^{\pi} A(\omega) |\mathcal{F}\left[w(n)e(n)\right]|^2 d\omega, \tag{18}$$

where $\mathcal{F}[\cdot]$ denotes the Fourier transform, $A(\omega) \in \{x \in \mathbb{R} | x > 0\}$ is a perceptual weighting function, $w(n)$ is the analysis window, and $e(n) = y(n) - x(n)$ is the modeling error. When the weighting function is chosen as the reciprocal of the masking threshold, the resulting error spectrum will be shaped like the masking threshold. While this measure is a spectral one, it is still inherently based on waveform matching since it operates on the Fourier transform of the time domain error, meaning that pre-echos, for example, will not go unpunished by the measure. With respect to audibility, the actual distortion values for non-stationary segments should be interpreted with care. In practice the spectral weighting function $A(\omega)$ is a discrete function, as is the error spectrum, and the distortion (18) is calculated as a summation of point-wise multiplications in the frequency domain. This corresponds to a circular filtering in the time domain. Putting this into matrix-vector notation, we get [13]

$$D = \|\mathbf{HW}(\mathbf{y} - \mathbf{x})\|_2^2, \tag{19}$$

where $\mathbf{H}$ is an circular matrix constructed from the impulse response of the filter corresponding to $\sqrt{A(\omega)}$ and $\mathbf{W}$ is a diagonal weighting matrix containing the elements of the analysis window $w(n)$. Depending on the filter length, it may still be advantageous to implement the filtering operation in the frequency domain. For further details on this procedure, we refer to, e.g., [13]. Using the perceptual distortion allows us to minimize a perceptually more meaningful measure than the 2-norm. However, doing so makes the pseudo-inverse $\mathbf{B}^\dagger$, defined in (15), frequency and segment dependent, forcing it to be re-calculated for each frequency and segment. Experimentally, we have found that the use of the perceptual distortion measure is much more important when minimizing wrt. the frequency in (9) than when solving for the AM coefficients in (14) and the phase in (13). Minimizing the perceptual distortion measure in (9) leads to the selection of the perceptually most important sinusoids. Thus, in order to minimize the complexity, we only apply the perceptual distortion measure in (9).

4 Audio Coding using the Decomposition

Many audio segments are well-modeled using a CA sinusoidal model, and applying the proposed AM decomposition is not always preferable from a rate-distortion point of view. Rather, to enable efficient coding of both stationary and transient segments, we propose the use of combined coder, containing both a CA sinusoidal coder and a coder based on the AM decomposition. Herein, the AM decomposition has been incorporated into the experimental coder described in [6]. Based on rate-distortion optimization, it is determined in each segment whether an AM or CA sinusoidal model should be used. We refer to such a combined coder as the AM/CA coder, using the term CA coder for the pure CA-based coder. Let $\mathcal{T}_s$ be a finite, discrete set of coding templates for segment s and $R(\tau)$ and $D(\tau)$ be the rate and distortion associated with coding template τ. Then, the problem of rate-distortion optimization under rate constraint (i.e., finding

the optimum distribution of $R^\star$ bits over S segments) can be written as the following unconstrained problem (see [2, 14] for further details)

$$\sum_{s=1}^{S} \min_{\tau \in \mathcal{T}_s} \left[D(\tau) + \lambda R(\tau) \right], \tag{20}$$

with $\lambda \geq 0$. This follows from the assumption that the (nonnegative) distortions and rates are independent and additive over the segments s. This means that the cost function can be minimized independently for each segment, for a given λ. Here we use the coding templates $\mathcal{T}_s = \{\psi_1, \ldots, \psi_{L_\psi}, \chi_1, \ldots, \chi_{L_\chi}\}$ with ψ_k being k constant-amplitude sinusoids and χ_k being k amplitude modulated sinusoids for segment s. When the optimal λ that leads to the target bit-rate $R^\star$, denoted $\lambda^\star$, has been found, the rate-distortion optimization simply becomes a matter of choosing the optimum coding template as

$$\tau_s^\star = \arg\min_{\tau \in \mathcal{T}_s} \left[D(\tau) + \lambda^\star R(\tau) \right]. \tag{21}$$

The optimal λ is found by maximizing the concave Lagrange dual function:

$$\lambda^\star = \arg\max_{\lambda} \left(\sum_{s=1}^{S} \left[\min_{\tau \in \mathcal{T}_s} D(\tau) + \lambda R(\tau) \right] - \lambda R^\star \right). \tag{22}$$

Typically, this is done by sweeping over λ (using some fast method exploiting the convexity of $R(D)$) until the rate $R(\lambda)$ is within some range of the target bit-rate [2]. We then chose between AM and CA using the following criterion

$$\min_k \left[D(\chi_k) + \lambda^\star R(\chi_k) \right] < \min_k \left[D(\psi_k) + \lambda^\star R(\psi_k) \right]. \tag{23}$$

Thus, AM coding template χ_k is chosen when it is the rate-distortion optimal choice among $\mathcal{T}_s$ for a particular segment.

5 Experimental Results

5.1 Configuration

In the experiments to follow, von Hann windows of length 30 ms were used in both analysis and overlap-add synthesis with 50% overlap. Sinusoidal parameters are quantized as follows: Phases are quantized uniformly using 5 bits/component, whereas amplitudes and frequencies are quantized in the logarithmic domain. Since entropy coding of the quantization indices is commonly used in audio coding, we estimate the resulting rates as the entropies of the quantization indices, which gives approximately 9 bits/component for frequencies and 6 bits/component for amplitudes. The AM coefficients are also quantized using the amplitude quantizer. This leads to an average of 30 bits/component

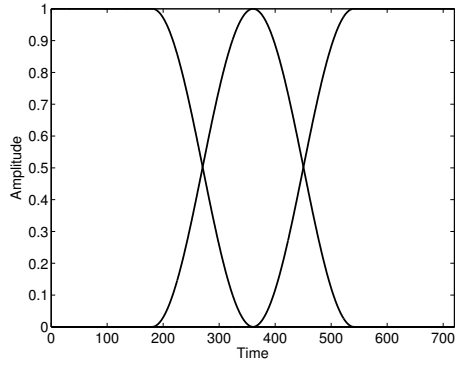

Figure 1: AM bases used in the experiments.

for amplitude modulated sinusoids and 20 bits/component for constant-amplitude. The quantizers were found to produce perceptually transparent results compared to original parameters. In the rate-distortion optimization, distortions are calculated using unquantized values as the measure (18) may be overly sensitive to frequency quantization. Note that the rates can be reduced significantly by differential encoding [15].

5.2 Informal Evaluation

Informal listening tests indicate that the combined AM/CA coder results in high perceived quality of coded excerpts for both stationary and transient parts. Generally, the type of signals that benefit from AM are signals that exhibit sharp onsets and stops, percussive sounds and changing signal types, such as transitions from unvoiced to voiced in speech signals. Often, the improvements are perceived as an increase in bandwidth. In Figure 2, the rate-distortion curves (or more correctly the distortion-rate curves) of the CA coder and the AM/CA coder are shown. These were found by sweeping over λ in (20) and finding the associated optimal rate and distortion point. It can be seen that there is a significant improvement in the rate-distortion tradeoff resulting from the proposed decomposition. It can also be seen that the curve saturates at higher rates, meaning that lower distortions cannot be achieved.

5.3 Listening Test

A blind AB preference test with reference was carried out on headphones using 6 different transient excerpts from SQAM with 7 inexperienced listeners participating. The listeners were asked to choose between the CA coder and the AM/CA coder, both operating at a bit-rate of approximately 30 kbps. Each experiment was repeated 8 times in a randomized, balanced way. The results are shown in Table 1. Significance was deter-

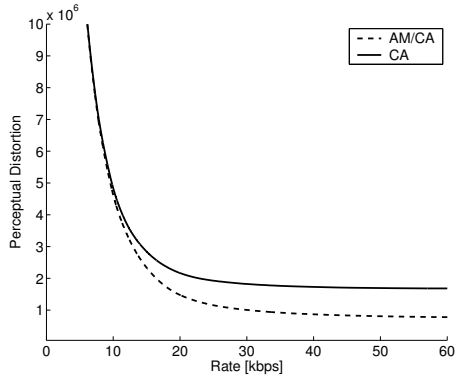

Figure 2: The rate-distortion curves of the CA coder (solid) and that of AM/CA coder (dashed) for for the excerpt Glockenspiel.

Results of Listening Tests			
	Preference [%]		
Excerpt	AM/CA	CA	Significant
Castanets	100	0	Yes
Claves	80	20	Yes
Glockenspiel	63	37	Yes
Harpsichord	63	37	Yes
Vibraphone	57	43	No
Xylophone	78	22	Yes
Total	74	26	Yes

Table 1: Results of AB-preference test.

mined using a binomial distribution and a one-sided test with a level of significance of 0.05. The test shows that performance can be improved significantly using the proposed decomposition.

6 Conclusion

In this paper, we have proposed a linear decomposition technique for amplitude modulated sinusoidal signals, showing that such a method might be used for high quality audio coding. Experiments indicate that a significantly higher rate of convergence, in terms of rate-distortion, can be achieved for transient segments when incorporating the proposed method in a combined coder. This is also confirmed by listening tests, show-

ing that for a given bit-rate, significant improvements can be gained for the coder using the proposed decomposition. These results are promising for applications of amplitude modulation in low bit-rate audio coding.

References

[1] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88(4), pp. 451–515, Apr. 2000.

[2] P. Prandoni, "Optimal Segmentation Techniques for Piecewise Stationary Signals," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne, 1999.

[3] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[4] M. G. Christensen, S. van de Par, S. H. Jensen, and S. V. Andersen, "Multiband amplitude modulated sinusoidal audio modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2004, pp. 169–172.

[5] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebart, "Advances in parametric coding for high-quality audio," in *114th Conv. Aud. Eng. Soc.*, 2003, paper preprint 5852.

[6] M. G. Christensen and S. van de Par, "Rate-distortion efficient amplitude modulated sinusoidal audio coding," in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2280–2284.

[7] G. Li, L. Qiu, and L. K. Ng, "Signal representation based on instantaneous amplitude models with application to speech synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 8(3), pp. 353–357, 2000.

[8] F. Myburg, A. C. den Brinker, and S. van Eijndhoven, "Sinusoidal analysis of audio with polynomial phase and amplitude," in *Proc. ProRISC*, 2001.

[9] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Processing*, vol. 47, pp. 2600–2603, Sept. 1999.

[10] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[11] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with application to target feature extraction," *IEEE Trans. Signal Processing*, vol. 44(2), pp. 281–295, Feb. 1996.

[12] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2001, pp. 1805 – 1808.

[13] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech and Audio Processing*, vol. 12(2), pp. 121–132, Mar. 2004.

[14] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1445–1453, Sept. 1988.

[15] J. Jensen and R. Heusdens, "A comparison of differential schemes for low-rate sinusoidal audio coding," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2003, pp. 205–208.

Paper F

On Perceptual Distortion Minimization and Nonlinear Least-Squares Frequency Estimation

Mads Græsbøll Christensen and Søren Holdt Jensen

The paper will appear in
IEEE Transactions on Audio, Speech and Language Processing, vol. 14(1),
pp. 99–109, January 2006.

The layout has been revised.

Abstract

In this paper, we present a framework for perceptual error minimization and sinusoidal frequency estimation based on a new perceptual distortion measure and we state its optimal solution. Using this framework, we relate a number of well-known practical methods for perceptual sinusoidal parameter estimation such as the pre-filtering method, the weighted matching pursuit and the perceptual matching pursuit. In particular, we derive and compare the sinusoidal estimation criteria used in these methods. We show that for the sinusoidal estimation problem, the pre-filtering method and the weighted matching pursuit are equivalent to the perceptual matching pursuit under certain conditions.

1 Introduction

The problem of estimating the parameters of a set of sinusoids in noise arises in many different applications. In digital processing of speech, the sinusoidal estimation problem arises in such applications as speech modeling and coding [1–5] and speech enhancement [6] and more recently, renewed interest in sinusoidal coding of speech has been spurred by the increasing interest in voice over packet-based networks [7–10]. Also in the field of audio processing, the sinusoidal signal model has been of interest for music analysis and synthesis [11–13], and parametric coding of audio [14–20]. In speech and audio processing the sinusoids can be seen as a parametric representation of the quasi-periodic, i.e. tonal, signal components, while the noise can be seen as the unvoiced, stochastic signal components [13]. The latter could, for example, be unvoiced speech, the bow noise of a violin, quantization errors or processing noise.

The applications mentioned above have in common that it is of interest to find a compact representation, or in other words to represent the signal in as few, physically meaningful parameters as possible. Since the end receiver of these signals is the human auditory system, it is also of interest to represent the perceptually most important components. In audio coding in particular, it is of interest to estimate and transmit only the parameters of audible sinusoids and in recent years, much effort has been put into this problem. Many different methods for solving this have been proposed, e.g. [21–28] all implement this in what seem to be different ways. Often, these methods rely heuristic rules taken from psychoacoustic experiments, while estimation theory, on the other hand, relies on statistical signal processing in finding model parameters. In [25] sinusoidal components are found in an iterative manner by assigning a perceptual weight to the spectrum and then picking the most dominant peak of the weighted spectrum. Another method is the so-called pre-filtering method, where the observed signal is filtered using a perceptual filter in order to achieve a weighting of the sinusoidal components, c.f. [26]. The methods of [27] and [28] are different methods yet—they rely on loudness and excitation pattern similarity criteria for sinusoidal component selection,

respectively.

In coding applications it is of particular interest to state the estimation criterion in a way that defines a distortion measure or metric. A globally optimal solution that minimizes this distortion measure ensures that at a given bit-rate (for a certain number of sinusoids in the case of sinusoidal coding), the lowest possible distortion is achieved. When the distortion measure is a perceptual one, meaning that it reflects the human auditory system, we can then claim that the perceived distortion is minimized at the given bit-rate. In linear predictive speech coding, for example, perception is traditionally taken into account using a fairly simple approach, where the noise spectrum is shaped by a perceptual weighting filter, which is derived directly from the linear prediction filter of the speech signal [29].

A recently published psychoacoustic masking model for audio coding has been shown to form a distortion measure [30, 31], and this distortion measure has been applied successfully to the sinusoidal estimation problem in [15, 23, 32, 33]. Based on this we define the perceptual frequency estimation problem and its optimal solution. We then analyze and relate a number of different practical perceptual frequency estimators that are all based on least-squares in this framework. In particular, we study the estimation criteria of these estimators. This allows us to analyze, quantify and understand the nature of the approximations made in these estimators. An important result is that the estimation criteria of the pre-filtering method and the weighted matching pursuit can be derived from the perceptual matching pursuit from the same assumption. Since many applications rely on a physical interpretation of the estimated parameters, the statistical properties of the estimators in question are also of significant importance. In that spirit we also investigate how the least-squares based estimators relate to estimation theory and maximum likelihood frequency estimation.

The rest of this paper is organized as follows. In Section 2 the frequency estimation problem is introduced along with the nonlinear least-squares frequency estimator. Then, in Section 3, we relate this to a simpler, common estimator, namely matching pursuit. In Section 4 we proceed to introduce a perceptual distortion measure that can be written in the form of a circulant, symmetric perceptual weighting matrix. In Section 5 we use this measure to formulate the perceptual frequency estimation problem and its optimal solution in terms of the perceptual nonlinear least-squares estimator. Moreover, we relate this to an approximation, namely the perceptual matching pursuit. The eigenvalue decomposition (EVD) of the perceptual weighting matrix and approximations with application to the problem at hand are studied in Section 6. In Section 7 we then show how this can be used to relate a number of well-known perceptual sinusoidal frequency estimators. We present some illustrative numerical examples in Section 8, and we summarize the results and give conclusions in Sections 9 and 10, respectively.

2 The Frequency Estimation Problem

The basic problem addressed in this paper can be stated as follows. Given a real observed signal $x(n)$ for $n = 0, \ldots, N-1$, find the parameters of the signal of interest $\hat{x}(n)$ in additive noise $e(n)$:

$$x(n) = \hat{x}(n) + e(n). \tag{1}$$

In our case the signal of interest $\hat{x}(n)$ is a sum of sinusoidal components

$$\hat{x}(n) = \sum_{l=1}^{L} A_l \cos\left(\omega_l n + \phi_l\right), \tag{2}$$

with each component having a constant amplitude A_l, initial phase ϕ_l, and frequency ω_l. The problem is then to estimate these parameters, in particular the frequencies $\boldsymbol{\omega} = [\, \omega_1 \ \cdots \ \omega_L \,]^T$. In the same process, the amplitudes and phases are usually also found, but as we shall see, these can written as complex linear parameters and can then be found in straightforward way.

Supposing that $e(n)$ is zero-mean white, i.i.d. (independent and identically distributed over observations) Gaussian noise of variance σ^2, the likelihood function $p(\mathbf{x}; \boldsymbol{\omega})$, which is a function of the observed signal and the model parameters (here only the frequencies) can be written as (see e.g. [34])

$$\begin{aligned} p(\mathbf{x}; \boldsymbol{\omega}) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left|x(n) - \hat{x}(n)\right|^2\right] \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left|x(n) - \hat{x}(n)\right|^2\right]. \end{aligned} \tag{3}$$

Introducing a vector containing the observed signal $\mathbf{x} = [\, x(0) \ \cdots \ x(N-1) \,]^T$ and a vector containing the modeled signal $\hat{\mathbf{x}} = [\, \hat{x}(0) \ \cdots \ \hat{x}(N-1) \,]^T$, this can be written as

$$p(\mathbf{x}; \boldsymbol{\omega}) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right]. \tag{4}$$

Taking the logarithm, we get the log-likelihood function

$$\ln p(\mathbf{x}; \boldsymbol{\omega}) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2. \tag{5}$$

We see that for white Gaussian noise, maximizing the likelihood function is the same as minimizing the squared error between the observed signal and the signal model.

In the nonlinear least-squares frequency estimator (NLS), the sinusoidal frequencies are estimated by minimizing exactly this error in a least-squares sense. The method is known as nonlinear least-squares as the cost function is nonlinear in the unknown frequencies. It is interesting, but perhaps not surprising, that in this particular case, the statistical approach of maximum likelihood (ML) turns into a deterministic method that matches the signal model to the outcome of the random process. The resulting estimator can be stated as the solution to the following problem [35]:

$$\min \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \min \|\mathbf{x} - \mathbf{Z}\mathbf{a}\|_2^2. \tag{6}$$

Here, the matrix $\mathbf{Z} \in \mathbb{C}^{N \times 2L}$ ($N > 2L$) is a so-called Vandermonde matrix[1] defined as

$$\mathbf{Z} = \begin{bmatrix} z_1^0 & z_1^{-0} & \cdots & z_L^0 & z_L^{-0} \\ z_1^1 & z_1^{-1} & \cdots & z_L^1 & z_L^{-1} \\ \vdots & \vdots & & \vdots & \vdots \\ z_1^{N-1} & z_1^{-(N-1)} & \cdots & z_L^{N-1} & z_L^{-(N-1)} \end{bmatrix}, \tag{7}$$

where signal poles $z_l = \exp(j\omega_l)$ come in complex conjugate pairs. Assuming that the signal poles are distinct, the matrix has full rank. Furthermore, we have that $\mathbf{a} \in \mathbb{C}^{2L}$, $\mathbf{a} = [\, a_1 \; a_1^* \; \cdots \; a_L \; a_L^* \,]^T$ with

$$a_l = \frac{A_l}{2} \exp(j\phi_l). \tag{8}$$

The NLS frequency estimates are then the combination of L frequencies (with $\hat{\cdot}$ denoting estimates) that minimizes the squared error, i.e.,

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \|(\mathbf{x} - \mathbf{Z}\mathbf{a})\|_2^2. \tag{9}$$

This can be formulated as a maximization problem using the principle of orthogonality:

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \mathbf{x}^H\mathbf{x} - \mathbf{x}^H\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x} \tag{10}$$

$$= \arg\max_{\boldsymbol{\omega}} \mathbf{x}^H\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}. \tag{11}$$

The corresponding amplitude and phase estimates are the solution to (6) given the frequencies:

$$\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}. \tag{12}$$

For more on estimation of amplitudes and phases, we refer the reader to the study in [37]. In order to solve the frequency estimation problem this way, we have to search

[1] Vandermonde matrices are sometimes defined to be square [36].

(numerically) for the combination of the L complex sinusoids that minimize the 2-norm of the error signal. This is essentially the subspace pursuit of [38] with the sum of sinusoids being the target subspace. Clearly, this is a complex procedure and it is not easily solved. In most real-time applications, solving this problem directly is not feasible. For more on the intractability of this problem, we refer the reader to [39].

One may argue that this point of view is unrealistic both in terms of solving the problem optimally and in terms of the assumptions with respect to the noise, but the NLS frequency estimator is very interesting from a theoretical point of view because it has excellent statistical performance. For the white Gaussian noise case, it is efficient and unbiased—it attains the Cramér-Rao Bound (see e.g. [35, 40, 41]).

In speech and audio processing the noise cannot generally be assumed to be white. For the colored noise case, with the Gaussian noise $e(n)$ now having the positive definite (non-diagonal) covariance matrix $\mathbf{\Sigma}$, the likelihood function is [41]

$$p(\mathbf{x};\boldsymbol{\omega}) = Q \exp\left[-\frac{1}{2}\left(\mathbf{x}-\hat{\mathbf{x}}\right)^H \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\hat{\mathbf{x}}\right)\right], \tag{13}$$

with

$$Q = \frac{1}{(2\pi)^{\frac{N}{2}}\sqrt{\det(\mathbf{\Sigma})}}. \tag{14}$$

The corresponding maximum likelihood estimator is then

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \left(\mathbf{x}-\hat{\mathbf{x}}\right)^H \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\hat{\mathbf{x}}\right). \tag{15}$$

Without prior knowledge of the noise covariance matrix $\mathbf{\Sigma}$, this problem is clearly more difficult to solve than for white noise where $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ and $\det(\mathbf{\Sigma}) = \sigma^{2N}$. However, as shown in [41], the NLS estimator in (11) is also asymptotically efficient for colored noise under some mild conditions. For more details on the relation between the NLS and ML estimators for the colored noise case and the associated Cramér-Rao bound, we refer the reader to [41], and for a practical method that achieves the Cramér-Rao bound see [42]. For non-Gaussian noise, the NLS estimator loses its maximum likelihood interpretation [41]. Here it must be stressed that we are not arguing as to the nature of noise in audio signals but rather as to the optimality of some commonly used methods that are based on least-squares.

3 Relaxation of the NLS Estimator

In this section we treat the relationship between the NLS frequency estimator and a well-known method for sinusoidal parameter estimation, namely matching pursuit [43]. As we shall see, there is a close relation between the two, although originally proposed in two entirely different contexts.

In matching pursuit a signal model is built iteratively by solving for one component at a time. This is done by finding the component from a dictionary, in this case composed of a set of complex sinusoids of different frequencies, that minimizes some norm (here the 2-norm) of the residual, which is formed by subtracting the i-th component from the i-th residual, i.e.,

$$r_{i+1}(n) = r_i(n) - \hat{A}_i \cos(\hat{\omega}_i n + \hat{\phi}_i), \tag{16}$$

with the residual being initialized as $r_1(n) = x(n)$. The Vandermonde matrix $\mathbf{Z}$ now contains the vector $\mathbf{z} = \left[\begin{array}{ccc} \exp(j\omega 0) & \cdots & \exp(j\omega(N-1)) \end{array} \right]^T$ and its complex-conjugate:

$$\mathbf{Z} = \left[\begin{array}{cc} \mathbf{z} & \mathbf{z}^* \end{array} \right]. \tag{17}$$

The frequency is then estimated as the minimizer of the 2-norm of the residual at iteration $i+1$

$$\hat{\omega}_i = \arg\min_{\omega} \|\mathbf{r}_{i+1}\|_2^2 = \arg\min_{\omega} \|\mathbf{r}_i - \mathbf{Z}\mathbf{a}\|_2^2 \tag{18}$$

$$= \arg\max_{\omega} \mathbf{r}_i^H \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{r}_i, \tag{19}$$

where $\mathbf{r}_i = \left[\begin{array}{ccc} r_i(0) & \cdots & r_i(N-1) \end{array} \right]^T$. After i iterations, the signal model is simply

$$\hat{x}_i(n) = \sum_{l=1}^{i} \hat{A}_l \cos(\hat{\omega}_l n + \hat{\phi}_l). \tag{20}$$

Writing out the estimation criterion (19) (here denoted J), we get

$$J = \mathbf{r}_i^H \mathbf{Z} \left(\mathbf{Z}^H \mathbf{Z}\right)^{-1} \mathbf{Z}^H \mathbf{r}_i \tag{21}$$

$$= \mathbf{r}_i^H \left[\begin{array}{cc} \mathbf{z} & \mathbf{z}^* \end{array} \right] \left[\begin{array}{cc} \mathbf{z}^H\mathbf{z} & \mathbf{z}^H\mathbf{z}^* \\ \mathbf{z}^T\mathbf{z} & \mathbf{z}^H\mathbf{z} \end{array} \right]^{-1} \left[\begin{array}{c} \mathbf{z}^H \\ \mathbf{z}^T \end{array} \right] \mathbf{r}_i. \tag{22}$$

We see that this is still a subspace pursuit, but in this case the subspace is a function of one variable ω. This is sometimes referred to as a conjugate-subspace pursuit [38]. Assuming that the complex sinusoid and its complex-conjugate are well separated in frequency (not close to 0 or π relative to N), the inner product between the two can be assumed to be zero[2]:

$$\mathbf{z}^H \mathbf{z}^* \approx 0. \tag{23}$$

The estimation criterion (22) can then be reduced significantly:

$$J = \mathbf{r}_i^H \left[\begin{array}{cc} \mathbf{z} & \mathbf{z}^* \end{array} \right] \left[\begin{array}{cc} \mathbf{z}^H\mathbf{z} & 0 \\ 0 & \mathbf{z}^H\mathbf{z} \end{array} \right]^{-1} \left[\begin{array}{c} \mathbf{z}^H \\ \mathbf{z}^T \end{array} \right] \mathbf{r}_i \tag{24}$$

$$= 2\frac{\left|\mathbf{z}^H \mathbf{r}_i\right|^2}{\mathbf{z}^H \mathbf{z}}. \tag{25}$$

[2]For the 2-norm case considered here, the conjugate-subspace pursuit can be solved efficiently without this assumption. However, this is not the case for the methods considered later in this paper.

The sinusoidal frequency estimation criterion can now be written in the well-known form

$$\hat{\omega}_i = \arg\max_{\omega} \frac{|\langle \mathbf{z}, \mathbf{r}_i \rangle|^2}{N}, \tag{26}$$

with $< \cdot, \cdot >$ denoting the inner product. The associated optimum complex scaling is

$$\hat{a}_i = \frac{\langle \mathbf{z}, \mathbf{r}_i \rangle}{N}, \tag{27}$$

which relates to the amplitude and phase in (16) as described in (8). We see that for the case of a sinusoidal dictionary MP is the NLS estimator in the one sinusoid case. It can be solved efficiently since the inner products $\langle \mathbf{z}, \mathbf{r}_i \rangle$ can be found using FFTs. Clearly, matching pursuit is a simplified approximation to (11). It can be seen as a relaxation of the original problem, where instead of solving the multidimensional nonlinear problem, we break it into several one-dimensional minimizations that have efficient implementations. Matching pursuit converges in the respective norm as i grows and the distortion is a non-increasing function of i (see [43]). It does not, generally, converge to zero in a finite number of iterations for the sinusoidal case as later iterations may introduce new spectral components due to the non-orthogonality of the components of redundant dictionaries. Sometimes this is also referred to as the readmission problem [44]. There are several ways to compensate for these problems (see for example [39, 44–47]).

On a historical note, the estimation procedure of [5, 11] first introduced in [48] is similar to that of matching pursuit for complex sinusoids later introduced in [43]. The RELAX algorithm [42] is an iterative sinusoidal frequency estimation algorithm, where the efficient solution to the one-sinusoid estimation problem is exploited in a recursive manner. It has been demonstrated to have excellent statistical performance achieving the Cramér-Rao bound for both white and colored Gaussian noise [41].

4 A Perceptual Distortion Measure

It is well-known that the 2-norm error measure does not correlate well with human sound perception. The choice of a distortion measure involves a trade-off between many factors. On one hand we would like to have a measure that takes as much of the processing in the human auditory into account as possible, while on the other hand we would like to have a measure which defines a mathematical norm. Another desirable property of the measure is that it can be incorporated in an efficient algorithm. A generalized perceptually weighted 2-norm can be written as

$$\|\mathbf{W}(\mathbf{x} - \hat{\mathbf{x}})\|_2^2, \tag{28}$$

where $\mathbf{W}$ is a so-called perceptual weighting or sensitivity matrix (e.g. [25, 49]). Even very sophisticated distortion measures can be expressed this way. For example, in [49]

the model of [50, 51] is linearized and put into the form of (28). Since we are here concerned with the estimation of stationary sinusoids, we assume the observed signal to be stationary. For stationary signals, the masking in the human auditory system is predominantly caused by simultaneous masking. Masking analysis in audio coding usually only considers distortions in the individual auditory filters, see e.g. ISO 11172-3 (MPEG-1) Psychoacoustic Model 1 described in [52]. Recently, it has been shown that significant improvements are gained by taking spectral integration into account [30, 31]. Using the masking model proposed in [30, 31], which was derived specifically for sinusoidal coding, the distortion D for a particular segment can be written as

$$D = \int_{-\pi}^{\pi} A(\omega)|E(\omega)|^2 d\omega, \tag{29}$$

where $A(\omega)$ is a real, positive perceptual weighting function and $E(\omega)$ is the discrete-time Fourier transform of the error $e(n) = w(n)\,[x(n) - \hat{x}(n)]$ where $w(n)$ is the analysis window. When the weighting function is chosen as the reciprocal of the masking threshold, the error spectrum which results from minimizing D will be shaped like the masking threshold.

In the coming analyses, we assume a rectangular window ($w(n) = 1\ \forall n$) for simplicity and mathematical convenience since we shall rely on asymptotic properties. In practice, the weighting function $A(\omega)$ and the error spectrum $E(\omega)$ are uniformly sampled spectra $A(k)$ and $E(k)$, respectively, and the integral (29) can be calculated as a summation of point-wise multiplications in the frequency domain:

$$D = \sum_{k=0}^{K-1} |\sqrt{A(k)}E(k)|^2. \tag{30}$$

The point-wise spectral multiplication corresponds to circular convolution in the time-domain, i.e.

$$\sum_{m=0}^{K-1} h(m)e((k-m)\,(\mathrm{mod}\,K)) \leftrightarrow \sqrt{A(k)}E(k), \tag{31}$$

with $\leftrightarrow$ denoting Fourier transform pairs. Furthermore, from Parseval's theorem, we have that the inner product can be calculated in the frequency domain as

$$\sum_{n=0}^{K-1} x^*(n)y(n) = \frac{1}{K}\sum_{m=0}^{K-1} X^*(k)Y(k). \tag{32}$$

This means that the discrete distortion measure (30) can be written as the 2-norm of a circular convolution:

$$D = \sum_{k=0}^{K-1} |\sum_{m=0}^{K-1} h(m)e((k-m)\,(\mathrm{mod}\,K))|^2. \tag{33}$$

The sampling frequency of the reciprocal of the masking curve $A(k)$ (and thus the length of the corresponding filter) is determined by the human auditory system and not by the input signal.

The distortion measure can now be put into the more convenient matrix-vector notation:

$$D = \|\mathbf{He}\|_2^2 \tag{34}$$

with $\mathbf{H}$ being the perceptual weighting matrix, in this case a filtering matrix, having the following structure

$$\mathbf{H} = \begin{bmatrix} h(0) & h(K-1) & \cdots & h(1) \\ h(1) & h(0) & \cdots & h(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ h(K-1) & h(K-2) & \cdots & h(0) \end{bmatrix}, \tag{35}$$

and $\mathbf{e} = [\, e(0) \ \cdots \ e(K-1) \,]^T$. This means that there is a duality between the spectral distortion measure and the two-norm of the circularly filtered error signal. This interpretation offers insights into the relation between a number of methods for perceptual frequency estimation. We will return to this later in the paper.

We now discuss how to derive an appropriate filter from the perceptual weighting function $A(k)$. As the perceptual filter has to be derived for each segment, computational complexity is of considerable importance. The simplest solution is to compute the impulse response as the inverse Fourier transform of $\sqrt{A(k)}$ for $n = 0, \ldots, K-1$, i.e.,

$$h(n) = \frac{1}{K} \sum_{k=0}^{K-1} \sqrt{A(k)} \exp\left(j2\pi kn/K\right) \tag{36}$$

$$= \frac{1}{K} \sum_{k=0}^{K-1} \sqrt{A(k)} \cos\left(2\pi kn/K\right), \tag{37}$$

where the last line follows from $A(k)$ being real and symmetric ($A(k) = A(K-k)$), which also means that $h(n)$ is symmetric, i.e. $h(n) = h(K-n)$. This procedure leaves us with an impulse response of length K while our observed signal is of length N. Typically, the required length of the spectral weighting function is higher than the number of time-samples, i.e. $N < K$. The signal and model vectors can then easily be zero-padded to length K or the last $K-N$ columns of $\mathbf{H}$ can be truncated. Filters of arbitrary order can be obtained using standard methods, and in the following sections we assume that the impulse response has been derived such that it has length N.

5 Perceptual NLS and MP

In many applications such as audio modeling and coding, it is of interest to extract only the perceptually most relevant sinusoidal component of the observed signal. Indeed, in audio coding, where the problem can be stated as minimizing the perceived distortion given some rate constraint, convergence in the perceptual distortion as we increase the number of sinusoids (and thus the rate) is desirable. Using the definitions in Section 4, we can restate the NLS frequency estimator as the following perceptually meaningful least-squares problem

$$\min \|\mathbf{H}(\mathbf{x} - \hat{\mathbf{x}})\|_2^2. \tag{38}$$

Let $\boldsymbol{\omega} = [\ \omega_1\ \cdots\ \omega_L\ ]^T$ be the set of frequencies that describe the Vandermonde matrix $\mathbf{Z} \in \mathbb{C}^{N\times 2L}$. Then the perceptual NLS estimates of the frequencies (and the corresponding optimal amplitudes and phases) are the solution to the problem

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \|\mathbf{H}(\mathbf{x} - \mathbf{Z}\mathbf{a})\|_2^2. \tag{39}$$

The vector $\hat{\boldsymbol{\omega}}$ is the vector containing the set of the frequencies of L sinusoids that minimize the filtered, weighted 2-norm and the vector $\hat{\mathbf{a}}$ contains the amplitudes and phases of those sinusoids in polar form. Since the filtering matrix is real and symmetric, i.e. $\mathbf{H}^H\mathbf{H} = \mathbf{H}^2$, these can be estimated as

$$\hat{\mathbf{a}} = \left(\mathbf{Z}^H\mathbf{H}^2\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{H}^2\mathbf{x}. \tag{40}$$

Substituting this into (39), we get

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \|\mathbf{H}(\mathbf{x} - \mathbf{Z}\mathbf{a})\|_2^2 \tag{41}$$

$$= \arg\max_{\boldsymbol{\omega}} \mathbf{x}^H\mathbf{H}^2\mathbf{Z}\left(\mathbf{Z}^H\mathbf{H}^2\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{H}^2\mathbf{x}. \tag{42}$$

This re-statement of the NLS frequency estimator allows us to estimate only the perceptually significant sinusoids and disregard inaudible ones, and to find the amplitudes and phases in such a way that artifacts are not introduced in the decoded signal. This formulation is only relevant when we are interested in a subset of the total number of sinusoids. Otherwise, there is no need for the spectral weighting of the error in the frequency estimation. However, the total number of sinusoids is generally unknown and robustness with respect to the number of sinusoids is desirable. We mention in passing that it also may be advantageous to incorporate the perceptual distortion in the estimation of amplitudes and phases as in (40) since erroneous estimates may introduce components in parts of the spectrum where no masker is present.

In terms of projections and transformations, the filtering matrix $\mathbf{H}$ can be thought of as a transformation to a perceptual domain and the problem of finding the optimal

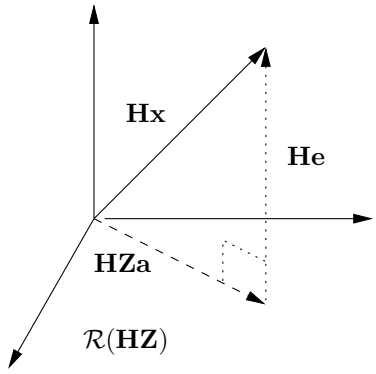

Figure 1: Orthogonal projection of the filtered input onto the column space of the filtered signal model.

signal model can be seen as a projection problem. Then, the transformed input signal is projected orthogonally onto the column space of the transformed signal model. This introduces an error which is orthogonal to the signal model in the perceptual domain. This is illustrated in Figure 1 with $\mathcal{R}(\cdot)$ denoting the range.

In the perceptual matching pursuit [23], which is a special case of the psychoacoustic adaptive matching pursuit with no adaptive norm, the dictionary element that minimizes the perceptual norm of the residual $\mathbf{r}_i$ is chosen. As in Section 3, this is just the one-sinusoid nonlinear least-squares estimator operating on the residual. The matrix $\mathbf{Z}$ again reduces to the vector $\mathbf{z} = \left[\begin{array}{ccc} \exp(j\omega 0) & \cdots & \exp(j\omega(N-1)) \end{array} \right]^T$, and the estimator is

$$\hat{\omega}_i = \arg\min_{\omega} \|\mathbf{H}(\mathbf{r}_i - \mathbf{z}a)\|_2^2. \tag{43}$$

with $\mathbf{r}_i$ again being the residual at iteration i (see section 3). Rewriting (43), we get the frequency estimator

$$\hat{\omega}_i = \arg\max_{\omega} \mathbf{r}_i^H \mathbf{H}^2 \mathbf{z} \left(\mathbf{z}^H \mathbf{H}^2 \mathbf{z} \right)^{-1} \mathbf{z}^H \mathbf{H}^2 \mathbf{r}_i \tag{44}$$

$$= \arg\max_{\omega} \frac{| \langle \mathbf{Hz}, \mathbf{Hr}_i \rangle |^2}{\|\mathbf{Hz}\|_2^2}, \tag{45}$$

and the associated optimal scaling, i.e. amplitude and phase, is

$$\hat{a}_i = \frac{\langle \mathbf{Hz}, \mathbf{Hr}_i \rangle}{\|\mathbf{Hz}\|_2^2}. \tag{46}$$

The perceptual MP converges in the perceptual distortion measure rather than the 2-norm. We see that as with matching pursuit and the one-sinusoid NLS estimator, there is an equivalence between the perceptual matching pursuit and the perceptual NLS. The perceptual MP can be implemented efficiently using two FFTs in each iteration.

6 EVD of the Perceptual Weighting Matrix

6.1 Signal Model Assumption

We now consider the example of a signal model component being an eigenvector $\mathbf{v}$ of the perceptual weighting matrix $\mathbf{H}$ with eigenvalue λ such that

$$\mathbf{H}\mathbf{v} = \lambda\mathbf{v}. \tag{47}$$

As we shall see in Section 7, this assumption leads to some interesting results and is indeed valid for certain important cases. It is well-known that complex sinusoids are eigenvectors of convolution operators, i.e.

$$\mathbf{v} = [\, \exp(j\omega 0) \;\cdots\; \exp(j\omega(N-1)) \,]^T. \tag{48}$$

For notational simplicity, we omit the dependence of the eigenvalue λ on the frequency ω. Strictly speaking, (47) holds only in general (i.e. for any ω) for the asymptotic case $N \to \infty$. For the following analysis, consider (47) to be simply an approximation.

The above simplification requires the calculation of eigenvalues for the different eigenvector approximations. The optimal approximation of the eigenvalue for the vector $\mathbf{v}$ in a least-squares sense can be stated as

$$\hat{\lambda} = \arg\min_{\lambda} \|\mathbf{H}\mathbf{v} - \lambda\mathbf{v}\|_2^2, \tag{49}$$

which is the Rayleigh coefficient, i.e.,

$$\hat{\lambda} = \frac{\mathbf{v}^H\mathbf{H}\mathbf{v}}{\mathbf{v}^H\mathbf{v}}. \tag{50}$$

We see that when the vector $\mathbf{v}$ is in fact an eigenvector of $\mathbf{H}$, this will result in the correct eigenvalue. The goodness of the eigenvalue approximation can conveniently be measured as

$$\|\mathbf{H}\mathbf{v} - \hat{\lambda}\mathbf{v}\|_2^2. \tag{51}$$

6.2 EVD of Circulant Matrices

In Section 6.1 we considered the assumption that the signal model components are eigenvectors of the filtering matrix. Now we take a look at the eigenvalue decomposition of circulant matrices, i.e. the filtering matrix $\mathbf{H}$, which is also symmetric. A circulant matrix, say $\mathbf{C} \in \mathbb{R}^{M\times M}$, has the following structure

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{M-1} & \cdots & c_1 \\ c_1 & c_0 & \cdots & c_{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M-1} & c_{M-2} & \cdots & c_0 \end{bmatrix}, \tag{52}$$

which is uniquely defined by the vector $\mathbf{c} = [\, c_0 \; \cdots \; c_{M-1} \,]^T$. Defining the discrete Fourier transform (DFT) matrix as

$$\mathbf{F} = \frac{1}{\sqrt{M}} \left[\begin{array}{cccc} \mathbf{f}_0 & \mathbf{f}_1 & \cdots & \mathbf{f}_{M-1} \end{array} \right], \tag{53}$$

with the individual Fourier bases $\mathbf{f}_k = \left[\begin{array}{ccc} f_k^0 & \cdots & f_k^{M-1} \end{array} \right]^T$ being composed from $f_k = \exp(j2\pi k/M)$. It then follows that the eigenvalue decomposition of the matrix $\mathbf{C}$ can be written as [36]

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \tag{54}$$

with $\mathbf{U} = \mathbf{F}^H$ and $\mathbf{\Lambda} = \sqrt{M}\,\mathrm{diag}(\mathbf{F}\mathbf{c})$. We see that the eigenvalues in the diagonal matrix $\mathbf{\Lambda}$ are simply the DFT coefficients of $\mathbf{c}$ and the eigenvectors contained in $\mathbf{U}$ are the Fourier bases of a DFT. For the special case of a symmetric $\mathbf{c}$, i.e. $c_m = c_{M-m}$, the eigenvalues are real.

6.3 Equivalent Forms

We now use the EVD to write the perceptual distortion measure in a number of different but equivalent forms. First, we write the perceptual distortion as

$$D = \|\mathbf{H}(\mathbf{x} - \hat{\mathbf{x}})\|_2^2 = (\mathbf{x} - \hat{\mathbf{x}})^H \mathbf{H}^2 (\mathbf{x} - \hat{\mathbf{x}}), \tag{55}$$

where $\mathbf{H}^2$ is also symmetric and circulant and has the eigenvalue decomposition $\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^H$. Here it is also interesting to note that comparing (55) to (15), we see that there is an inherent contradiction in the use of the perceptual weighting matrix and the inverse covariance matrix in the maximum likelihood estimator for the colored noise case since $\mathbf{H}^2 \neq \mathbf{\Sigma}^{-1}$. Now the perceptual weighting can be rewritten into the following diagonal form:

$$\begin{aligned} D &= (\mathbf{x} - \hat{\mathbf{x}})^H \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^H (\mathbf{x} - \hat{\mathbf{x}}) && (56) \\ &= (\mathbf{U}^H\mathbf{x} - \mathbf{U}^H\hat{\mathbf{x}})^H \mathbf{\Lambda}^2 (\mathbf{U}^H\mathbf{x} - \mathbf{U}^H\hat{\mathbf{x}}). && (57) \end{aligned}$$

We note that the signal model $\hat{\mathbf{x}}$ may be chosen such that $\mathbf{U}^H\hat{\mathbf{x}}$ can be found analytically or pre-computed and stored in memory. Windowed sinusoids, for example, have simple Fourier transforms. As another example of this, we now treat the case of transform coding with the signal model components being equivalent to the eigenvectors, i.e. $\hat{\mathbf{x}} = \mathbf{U}\mathbf{y}$. In transform coding, the optimization problem concerns the transform coefficients $\mathbf{y}$. Bits are allocated such that the perceptual error is minimized. Now, the perceptual distortion can be rewritten as

$$D = (\mathbf{U}^H\mathbf{x} - \mathbf{y})^H \mathbf{\Lambda}^2 (\mathbf{U}^H\mathbf{x} - \mathbf{y}), \tag{58}$$

or the equivalent form where the input signal $\mathbf{x}$ is pre-filtered:

$$D = \|\mathbf{Hx} - \mathbf{HUy}\|_2^2 = \|\mathbf{Hx} - \mathbf{U\Lambda y}\|_2^2 . \tag{59}$$

It can be seen that distortion calculations can be simplified this way. This is a significant advantage in coding based on rate-distortion optimization [53], which requires the calculation of distortions for different allocations and quantizers.

7 Relation to Simplified Estimators

7.1 Pre-filtering Method

Using the eigenvector assumption in (47) the sinusoidal frequency estimation criterion (38) can be significantly simplified:

$$\min \|\mathbf{H}(\mathbf{r}_i - \hat{\mathbf{r}}_i)\|_2^2 = \min \|\mathbf{H}\left(\mathbf{r}_i - \mathbf{v}a\right)\|_2^2 \tag{60}$$

$$= \min \|\mathbf{Hr}_i - \lambda \mathbf{v}a\|_2^2, \tag{61}$$

where a is a complex scale factor (amplitude and phase in polar form), which is included here since we do not restrict the norm or the phase of $\mathbf{v}$. The optimal value of this scale factor can then easily be found as

$$\hat{a} = \frac{\mathbf{v}^H \lambda^* \mathbf{Hr}_i}{\mathbf{v}^H |\lambda|^2 \mathbf{v}}. \tag{62}$$

Next, expressing the perceptual NLS in terms of the unknown eigenvector, the frequency estimation criterion is simplified significantly:

$$\hat{\omega}_i = \arg\min_{\omega} \|\mathbf{Hr}_i - \lambda \mathbf{v}a\|_2^2 \tag{63}$$

$$= \arg\max_{\omega} \frac{\mathbf{r}_i^H \mathbf{H}^H \lambda \mathbf{v}\mathbf{v}^H \lambda^* \mathbf{Hr}_i}{\mathbf{v}^H \lambda^* \lambda \mathbf{v}} \tag{64}$$

$$= \arg\max_{\omega} \frac{|\langle \mathbf{v}, \mathbf{Hr}_i \rangle|^2}{N}. \tag{65}$$

We see that the estimator reduces to maximizing the inner product between the eigenvector $\mathbf{v}$ and $\mathbf{r}_i$ filtered by the perceptual filter. This inner product is just the periodogram of the perceptually filtered observed signal since $\mathbf{v}$ is a complex sinusoid. The modification of the signal model due to the filtering cancels out in the selection criterion and can be ignored. This is, however, not the case for damped sinusoids and pre-filtering is not well justified in that case. In practice this means that the input has to be filtered by the perceptual filter and then a squared error measure may be minimized in the estimation procedure if the model component is an eigenvector of $\mathbf{H}$ or is a reasonable approximation thereof.

The pre-filtering method has been applied to the perceptual estimation problem in e.g. [26, 54].

7.2 Pre- and Post-filtering Method

In the pre- and post-filtering approach of [55, 56], modeling is performed in the perceptual domain, i.e. operating on the pre-filtered signal:

$$\min \|\mathbf{H}\mathbf{r}_i - \hat{\mathbf{p}}\|_2^2 = \min \|\mathbf{p} - \hat{\mathbf{p}}\|_2^2. \tag{66}$$

Afterward, the modeled signal $\hat{\mathbf{p}}$ has to be mapped back to the signal domain by the inverse filter (also called the post-filter)

$$\hat{\mathbf{r}}_i = \mathbf{H}^{-1}\hat{\mathbf{p}}, \tag{67}$$

which means that the post-filter has to be sent to the decoder in coding applications. Otherwise, this approach differs from the pre-filtering method in algorithmic form in that the signal model is modified after the estimation/quantization rather than before. This has the advantage that the structure of the model, which may be lost by the filtering, is preserved in the estimation/quantization process. However, to argue that the signal model $\hat{\mathbf{p}}$ should be posed in the perceptual domain rather than in the signal domain seems somewhat contrived as the physical meaning of the model parameters is potentially lost in the transformation.

If the signal model component $\hat{\mathbf{p}}$ is an eigenvector of the inverse perceptual filter $\mathbf{H}^{-1}$, the post-filtering can be reduced to a simple scaling,

$$\hat{\mathbf{r}}_i = \lambda\hat{\mathbf{p}}, \tag{68}$$

in which case the signal model is valid also in the perceptual domain and can be modified directly. Also, the post-filter does not have to be transmitted to the receiver in this case.

For some types of estimators, though, the pre-filtering of the input signal has some serious drawbacks. Since it colors the signal, any noise will also be colored. The performance of subspace-based estimators degrades when the noise is not white [35]. Typically, this would be solved by applying pre-whitening but that is not an option for this application. These arguments favor NLS-based approximations such as matching pursuit for perceptual frequency estimation since NLS is still asymptotically efficient for colored noise [41].

7.3 Weighted Matching Pursuit

Since the filtering matrix $\mathbf{H}$ is symmetric, i.e. $\mathbf{H}^H = \mathbf{H}$, the inner product in the numerator of (65) can be written as

$$\langle \mathbf{v}, \mathbf{H}\mathbf{r}_i \rangle = \mathbf{v}^H\mathbf{H}\mathbf{r}_i = (\lambda\mathbf{v})^H \mathbf{r}_i, \tag{69}$$

such that the component selection criterion becomes

$$\hat{\omega}_i = \arg\max_{\omega} \frac{|\langle \mathbf{v}, \mathbf{H}\mathbf{r}_i \rangle|^2}{N} = \arg\max_{\omega} |\lambda|^2 \frac{|\langle \mathbf{v}, \mathbf{r}_i \rangle|^2}{N}. \tag{70}$$

The perceptual filtering approach can thus be reduced to a simple weighting of the inner products, where the weight is the absolute value of the eigenvalue associated with the eigenvector $\mathbf{v}$. This is in fact what the weighted matching pursuit does [25]. In the weighted MP the eigenvalue of a sinusoid of frequency ω is approximated as

$$\hat{\lambda} \approx \sqrt{A\left(\left\lfloor \frac{\omega K}{2\pi} + \frac{1}{2} \right\rfloor\right)}, \tag{71}$$

rather than the computationally more demanding least-squares approximation in (50). We see from (70) that under certain conditions on the perceptual filter, the sinusoidal estimator weighted MP is identical to the pre-filtering method. In [25], the weighting is introduced as a heuristic for incorporating psychoacoustics. Here, we have established the method as an approximation of the perceptual NLS.

The weighted MP has the problem that due to the perceptual weighting, the selected components may not be spectral maxima and spectral distortion introduced by the side-lobes of the sinusoidal components are not taken into account. This may cause audible artifacts. In the perceptual MP these problems are solved, and listening tests in [23] demonstrated its superior performance. The problems of the weighted MP can though easily be fixed by adding the constraints that the estimates have to be spectral maxima.

8 Numerical Examples

In this section we illustrate some of the points made in the previous sections using an example of a sinusoidal audio signal, the trumpet signal of SQAM [57]. In Figure 2 a segment of this signal is shown. The signal is sampled at 44.1 kHz. The masking curve is derived using the model in [30] and the corresponding perceptual weighting function is shown in Figure 3 along with the periodogram of the segment in Figure 2. Note the very distinct peaks and the harmonic structure in the periodogram.

The convergence of the perceptual MP in the perceptual norm is illustrated in Figure 4, again for the trumpet signal in Figure 2. Note how the perceptual distortion is a non-increasing function of the number of components. The sinusoidal frequencies that are estimated in the individual iterations of the perceptual MP (indicated by numbers) are shown in Figure 5. The effect of the perceptual distortion measure can been observed in that although more energy is present at peak 2, the perceptual MP picks peak 1 first. From the figure it is clear that the effect of the perceptual distortion measure is one of ordering. In Figure 6 an illustration of the error introduced by the eigenvector/-value approximation is shown. The figure shows the perceptual weighting for a segment of

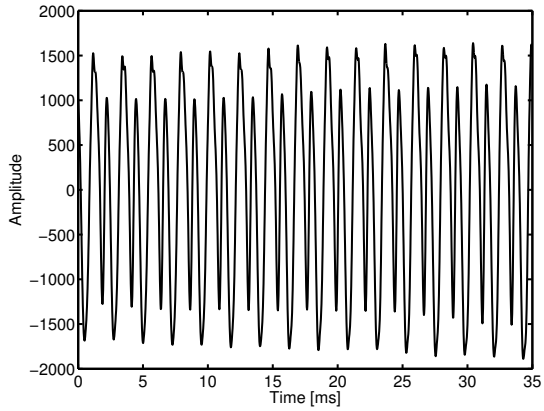

Figure 2: Example of an audio segment, trumpet. The trumpet signal is a fairly stationary, tonal signal.

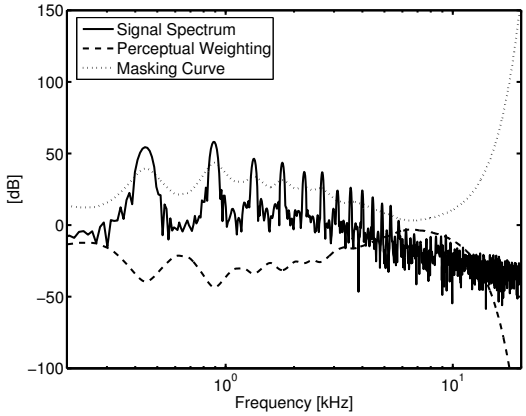

Figure 3: Perceptual weighting function (dashed), masking curve (dotted) and spectrum for the trumpet signal (solid) in Figure 2.

the trumpet signal and the error as defined by (51) introduced as a function of frequency with the eigenvalues being approximated using (50). Also shown is the signal-to-noise ratio (SNR), which is calculated as

$$SNR = 10\log_{10}\frac{\|\mathbf{Hv}\|_2^2}{\|\mathbf{Hv}-\hat{\lambda}\mathbf{v}\|_2^2}[\mathrm{dB}]. \tag{72}$$

The perceptual weighting was derived with a frequency resolution of 4096 uniformly spaced points, and the corresponding filter was calculated by taking the inverse discrete

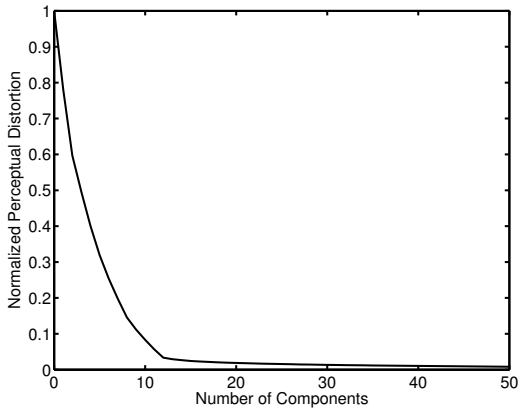

Figure 4: Convergence of the perceptual matching pursuit in the perceptual distortion for the trumpet signal in Figure 2.

Fourier transform of its square-root. The complex sinusoids were windowed by a Hanning window having a length of 1544 samples and then zero-padded to length 4096 to match the length of the perceptual filter. These are fairly typical choices of constants in audio coding. From this figure, it is very clear that these windowed, zero-padded complex sinusoids are not eigenvectors of the filtering matrix, since the SNR is far from the numerical noise floor (64 bit floating point). The loss in estimator performance in terms of distortion may well be worth it, though, as considerable complexity reductions can be achieved. It can also be seen that the goodness of the approximation is highly frequency dependent with the approximation performing well at high frequencies for this particular perceptual weighting function. This can be attributed to the perceptual weighting function being flatter in this region. Note that the perceptual weighting function will be dominated by the threshold in quiet for very low and high frequencies. When the length of the perceptual filter and the complex sinusoids are the same and no window is applied, the error hits the numerical noise floor as the complex sinusoids become eigenvectors of the filtering matrix.

9 Results

In this section we briefly summarize and discuss the main results of this paper. In particular we recapitulate the conditions under which the different methods that have been discussed are equivalent and optimal.

- When estimating the frequencies of sinusoids in additive white and Gaussian noise, the nonlinear least-squares method is the maximum likelihood estima-

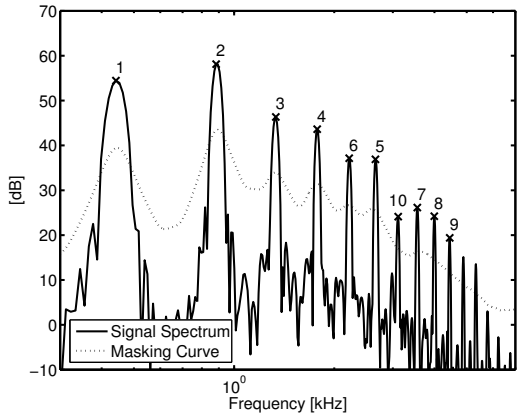

Figure 5: Frequencies estimated (crosses) in the individual iterations (indicated by number) by the perceptual matching pursuit.

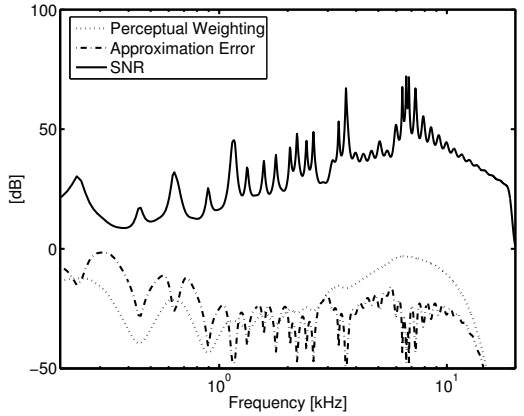

Figure 6: The error (solid) and SNR (dashed) as function of frequency introduced by the eigenvector approximation for a particular perceptual weighting function (dotted).

tor. The nonlinear least-squares method is efficient, i.e. in this case it attains the Cramér-Rao bound and is hence optimal.

- Under the condition that the noise is Gaussian but colored, there is an equivalence between maximum likelihood and (weighted) least-squares based estimation. The non-linear least-squares method is still asymptotically optimal in this case.

- The matching pursuit algorithm is a (one-dimensional) relaxation of the subspace pursuit of nonlinear least-squares. It converges in the respective norm, here the 2-norm, as a function of the number of components and has an efficient implementation for the frequency estimation problem.
- A recently established perceptual distortion measure, which shapes the error spectrum according to the masking threshold, can be shown to form a circulant and symmetric perceptual weighting matrix, which can be interpreted as a filtering matrix. Circulant and symmetric weighting matrices have eigenvectors that are rectangularly windowed complex sinusoids of uniformly spaced frequencies. Asymptotically, sinusoids of arbitrary frequencies are eigenvectors of the weighting matrix.
- When this perceptual weighting matrix is applied in solving the least-squares problems in the NLS and MP estimators, we get the perceptual nonlinear least-squares estimator and the simpler perceptual matching pursuit. The perceptual matching pursuit now converges in the perceptual distortion.
- The pre-filtering method and the weighted matching pursuit are equivalent to the perceptual matching pursuit when the model components are eigenvectors of the perceptual weighting matrix. This allows for very efficient implementation of the perceptual weighting. In some applications of the pre-filtering method and the weighted matching pursuit, the model components are not eigenvectors of the weighting matrix; then, these methods are only approximations of the perceptual matching pursuit.

10 Conclusion

We have introduced the perceptual frequency estimation problem based on a spectral distortion measure and its optimal solution, the nonlinear least-squares frequency estimator. The nonlinear least-squares method has a strong background in statistical signal processing and estimation theory and is well-known to have excellent statistical performance. We have related this to a number of well-known methods for perceptual parameter estimation, namely the perceptual matching pursuit, the weighted matching pursuit and the pre-filtering method. It has been shown that these methods can be seen as relaxations and approximations of the optimal solution. Specifically, we have established the perceptual matching pursuit as a relaxation of the nonlinear least-squares estimator, and we have shown that it reduces to the pre-filtering method and the weighted matching pursuit under certain conditions.

References

[1] P. Hedelin, "A tone oriented voice excited vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1981, pp. 205–208.

[2] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 4.

[3] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(4), pp. 744–754, Aug. 1986.

[4] R. J. McAulay and T. F. Quatieri, "Speech Transformation Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1449–1464, Dec. 1986.

[5] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech and Audio Processing*, vol. 5(5), pp. 389–406, Sept. 1997.

[6] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 731–740, Oct. 2001.

[7] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech and Audio Processing*, 2004, accepted.

[8] C. A. Rødbro, M. G. Christensen, S. H. Jensen, and S. V. Andersen, "Compressed domain packet loss concealment of sinusoidally coded speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2003, pp. 104–107.

[9] C. A. Rødbro, J. Jensen, and R. Heusdens, "Rate-distortion optimal time-segmentation and redundancy selection for VoIP," *IEEE Trans. Speech and Audio Processing*, 2005, accepted.

[10] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, "Hidden markov model based framework for packet loss concealment in voice over IP," *IEEE Trans. Speech and Audio Processing*, 2005, accepted.

[11] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis-synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40(6), pp. 497–516, 1992.

[12] J. O. Smith III and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," in *Proceedings of the International Computer Music Conference (ICMC-87, Tokyo), Computer Music Association*, 1987.

[13] J. O. Smith and X. Serra, "Spectral Modelling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, 1990.

[14] K. N. Hamdy, M. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 1045–1048.

[15] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[16] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC – Analysis/Synthesis Audio Codec for Very Low Bit Rates," in *100th Conv. Aud. Eng. Soc.*, 1996, paper preprint 4179.

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

[18] E. G. P. Schuijers, A. W. J. Oomen, A. C. den Brinker, and J. Breebart, "Advances in parametric coding for high-quality audio," in *114th Conv. Aud. Eng. Soc.*, 2003, paper preprint 5852.

[19] T. S. Verma and T. H. Y. Meng, "A 6kbps to 85kbps scalable audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 877–880.

[20] S. N. Levine and J. O. Smith III, "A switched parametric & transform audio coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 985–988.

[21] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. Audio Eng. Soc. 17th Conf: High Quality Audio Coding*, 1999, pp. 244–250.

[22] R. Vafin, S. V. Andersen, and W. B. Kleijn, "Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2000, pp. 901–904.

[23] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[24] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. van Huffel, "Perceptual audio modeling with exponentially damped sinusoids," *Signal Processing*, vol. 85, pp. 163–176, 2005.

[25] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1999, pp. 981–984.

[26] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech and Audio Processing*, vol. 12(2), pp. 121–132, Mar. 2004.

[27] H. Purnhagen, N. Meine, and B. Edler, "Sinusoidal coding using loudness-based component selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2002, pp. 1817–1820.

[28] T. Painter and A. S. Spanias, "Perceptual segmentation and component selection in compact sinusoidal representation of audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, 2001, pp. 3289–3292.

[29] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27(3), pp. 247–254, 1979.

[30] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2001, pp. 1805 – 1808.

[31] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2004.

[32] R. Heusdens, J. Jensen, W. B. Kleijn, V. kot, O. A. Niamut, S. van de Par, N. H. van Schijndel, and R. Vafin, "Bit-rate scalable intra-frame sinusoidal audio coding based on rate-distortion optimisation," *J. Audio Eng. Soc.*, 2005, submitted.

[33] M. G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(4), July 2006.

[34] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[35] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Prentice Hall, 1997.

[36] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[37] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Trans. Signal Processing*, vol. 48(2), pp. 338–352, Feb. 2000.

[38] M. M. Goodwin, "Matching pursuit and atomic signal models based on recursive filter banks," *IEEE Trans. Signal Processing*, vol. 47(7), pp. 1890–1902, July 1999.

[39] G. Davis, "Adaptive nonlinear approximations," Ph.D. dissertation, New York University, Sept. 1994.

[40] P. Stoica, R. L. Moses, B. Friedlander, and T. Söderström, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(3), pp. 378–392, Mar. 1989.

[41] P. Stoica, A. Jakobsson, and J. Li, "Cisiod parameter estimation in the coloured noise case: Asymptotic cramér-rao bound, maximum likelihood, and nonlinear least-squares," in *IEEE Trans. Signal Processing*, vol. 45(8), Aug. 1997, pp. 2048–2059.

[42] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with application to target feature extraction," *IEEE Trans. Signal Processing*, vol. 44(2), pp. 281–295, Feb. 1996.

[43] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41(12), pp. 3397–3415, Dec. 1993.

[44] M. M. Goodwin, "Adaptive Signal Models: Theory, Algorithms, and Audio Applications," Ph.D. dissertation, University of California, Berkeley, 1997.

[45] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with application to wavelet decomposition," in *Conference Records of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers*, vol. 1, 1993, pp. 40–44.

[46] H. Feichtinger, A. Turk, and T. Strohmer, "Hierarchical parallel matching pursuit," *Proc. of the SPIE - The International Society for Optical Engineering*, vol. 2302, pp. 222–232, 1994.

[47] J. Adler, B. Rao, and K. Kreutz-Delgado, "Comparison of basis selection methods," in *Conference Records of the Thirtieth Asilomar Conference on Signals, Systems, and Computers*, vol. 1, 1996, pp. 252–257.

[48] E. B. George and M. J. T. Smith, "A new speech coding model based on a least-squares sinusoidal representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1987, pp. 1641–1644.

[49] J. Plasberg, D. Zhao, and W. B. Kleijn, "Sensitivity matrix for a spectro-temporal auditory model," in *Proc. XII European Signal Processing Conf. (EUSIPCO)*, 2004, pp. 1673–1676.

[50] 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. Am.*, vol. 99(6), pp. 3615–3622, June 1996.

[51] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. ii. simulations and measurements," *J. Acoust. Soc. Am.*, vol. 99(6), pp. 3623–3631, June 1996.

[52] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88(4), pp. 451–515, Apr. 2000.

[53] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1445–1453, Sept. 1988.

[54] R. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2004, pp. 189–192.

[55] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 881–884.

[56] G. D. T. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual audio coding using adaptive pre- and post-filters and lossless compression," in *IEEE Trans. Speech and Audio Processing*, vol. 10(6), Sept. 2002, pp. 379–390.

[57] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.

Paper G

Open Loop Rate-Distortion Optimized Sound Coding

Fredrik Nordén, Mads Græsbøll Christensen, and Søren Holdt Jensen

The paper has been published in
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 161–164, 2005.

Abstract

This paper addresses complexity reduced rate-distortion optimized audio coding under rate constraint. A technique where distortion minimizing coding templates, chosen from a set of templates, are jointly selected for a set of segments. This optimization requires knowledge of rate-distortion pairs for all segments, and for each coding template, which often are costly to obtain. The proposed framework exchanges true rate-distortion pairs with predicted ones, thereby allowing for complexity reduction. The prediction is based on a property vector extracted for each segment, from which distortion predictions, using Gaussian mixture models, are performed. Here, we evaluate the proposed framework in a sinusoidal coding context. The results show that the proposed framework can increase the distortion performance, compared to a fixed sinusoidal coding scheme.

1 Introduction

Rate-distortion (R-D) optimization is of interest for audio coding for several reasons. It allows for adaptive coding schemes, where the coder is adapted to user and network constraint as well as source characteristics, thereby increasing the overall distortion performance. For example, parametric coders typically outperform transform coders at low bit-rates, and LPC-based coders perform very well for speech but not for audio. An R-D optimized selection among such a set of coders is thus of interest.

There are a multitude of different applications that can be put into the R-D optimization framework: 1) Coder selection for specific segments [1], 2) Distribution of bits over stages in multistage structures [2], 3) Variable bit-rate (optimal distribution of bits over segments) [3], and 4) Dynamic time-segmentation [3, 4]. All of these applications require knowledge of the incurred distortion in the current audio segment for all of the coders (coding template, number of sinusoids, etc), in order to perform R-D optimization. For some of the above applications, we end up having to do distortion calculations, which sometimes require both signal analysis and synthesis, for many different coding templates, not necessarily useful in the final coder synthesis.

The complexity of these distortion calculations may be severe, preventing the use of R-D optimized coders in many applications. Thus, we here propose an open loop approach to the R-D optimization problem. We exchange coding distortions with predicted ones, thereby allowing for complexity reduction. For the prediction purpose we employ an open loop framework for distortion prediction proposed in [5]. The framework is based on a property vector extracted from the segment to be coded, from which distortion predictions, using a Gaussian mixture model (GMM) of the joint property-distortion pdf, are performed. We evaluate the proposed framework in a sinusoidal coding context. Based on predicted R-D curves we perform R-D optimized distribution of sinusoids over sets of segments matching a given bit-budget. The results are compared with a sinusoidal coder optimized on original R-D curves, and a sinusoidal coder

using a fixed number of sinusoids per segment.

The paper is organized as follows. In Section 2 we discuss the basics of R-D optimized coding, and in Section 3 we present the prediction framework. This is followed by a presentation of the experimental setup in Section 4. In Section 5 we evaluate the goodness of the proposed system. Finally, we conclude in Section 6.

2 Rate-Distortion Optimization

The problem of distributing a certain number of bits over a set of segments, $\mathcal{S}$, constituting an optimization viewport, can be cast into rate-distortion optimization under rate constraint. This optimization can be stated as the following constrained optimization problem:

$$\begin{aligned} \min \quad & D \\ \text{s.t.} \quad & R \leq R^\star, \end{aligned} \tag{1}$$

where D is the distortion, R is the resulting rate, and $R^\star$ is the target rate. Let $\mathcal{T}_s$ be a finite, discrete set of coding templates (ways of encoding, etc.) for segment s, and $R(\tau)$ and $D(\tau)$ be the rate and distortion associated with coding template $\tau \in \mathcal{T}_s$. The distortion D and the rate R are the sum of distortions and rates over the segments, $\mathcal{S}$, associated with a particular set of coding templates $\boldsymbol{\tau} = [\tau_1 \ \cdots \ \tau_S]$ with $\tau_i \in \mathcal{T}_i$, i.e.,

$$D = \sum_{s=1}^{S} D(\tau_s) \quad \text{and} \quad R = \sum_{s=1}^{S} R(\tau_s). \tag{2}$$

The problem (1) can then be written as the following unconstrained problem [4]

$$\min_{\boldsymbol{\tau}} \sum_{s=1}^{S} D(\tau_s) + \lambda R(\tau_s) = \sum_{s=1}^{S} \min_{\tau \in \mathcal{T}_s} D(\tau) + \lambda R(\tau), \tag{3}$$

where λ is the non-negative Lagrange multiplier. The right side follows from assuming that distortions and rates are additive and independent over segments. This means that the optimization problem can be solved independently for each segment for a particular λ. The Lagrange multiplier λ can be interpreted as the slope of the R-D curve for a certain rate. The problem is then to find the $\lambda^\star$ that leads to the target bit rate $R^\star$. Such a λ cannot be guaranteed to exist for discrete problems such as ours. We can, however, find a solution close to the optimal one provided that the $\{R(\tau), D(\tau)\}$ points are sufficiently dense. The optimal λ is found by maximizing the concave Lagrange dual function:

$$\lambda^\star = \arg\max_{\lambda} \left[\sum_{s=1}^{S} \left(\min_{\tau \in \mathcal{T}_s} D(\tau) + \lambda R(\tau) \right) - \lambda R^\star \right]. \tag{4}$$

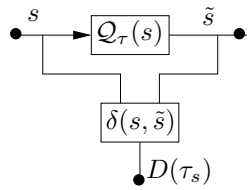

Figure 1: Illustration of the evaluation of the incurred distortion, $D(\tau_s)$, for one particular coding template, τ, and one particular audio segment, s. $\mathcal{Q}_\tau(\cdot)$ represents the coding or modeling associated with template, τ, and $\delta(\cdot)$ is the distortion criterion.

This can be done by sweeping over λ using simple bisection until the rate $R(\lambda)$ is within some range of the target bit rate [4]. Given the optimal $\lambda^\star$, the rate-distortion optimization simply becomes a matter of choosing the optimum coding template for a particular segment s as

$$\tau_s^\star = \underset{\tau \in \mathcal{T}_s}{\arg\min} \left[D(\tau) + \lambda^\star R(\tau) \right]. \tag{5}$$

For the rate-distortion optimization to result in improvements in perceived quality, the chosen distortion criterion, $\delta(\cdot)$, must reflect human sound perception. In this work we have chosen to work with the distortion criterion proposed in [6], which is further described in Section 4.

3 Rate-Distortion Prediction

To perform R-D optimized coding over a set of segments, $\mathcal{S}$, using a set of coding templates, $\mathcal{T}_s$, we require knowledge of R-D points for each segment and each coding template,

$$\{R(\tau_s), D(\tau_s)\} : \forall \ s \in \mathcal{S}, \ \forall \ \tau_s \in \mathcal{T}_s. \tag{6}$$

Ideally these points are found by coding each segment with each of the coding templates, as visualized in Figure 1[1]. This approach is highly complex, and in general therefore not feasible. Thus we here suggest an open loop alternative, where distortions, $\{D(\tau_s)\}$, are predicted from the current segment of audio, s, as visualized in Figure 2. In essence the structure in Figure 1 is exchanged for the structure in Figure 2. Below, we discuss the predictor employed to predict the incurred distortion for one particular coding template. In practice we require one predictor, as described below, for each coding template.

[1]The structure in Figure 1 needs to be processed NxM times, if we perform a joint optimization over N segments, using M coding templates for each segment.

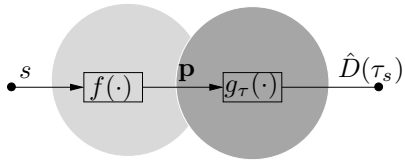

Figure 2: A framework for prediction of the incurred distortion, $D(\tau_s) = \delta(s, \mathcal{Q}_\tau(s))$, when coding a random vector s, using coding template τ. A dimension reducing property vector extraction, $f(\cdot)$, is followed by a distortion prediction, $g_\tau(\cdot)$.

3.1 Property Vector Based Prediction

We employ distortion prediction as suggested in [5]. The overall prediction is separated into a property extraction, $f(\cdot)$, and a prediction, $g_\tau(\cdot)$, as visualized in Figure 2. Each audio segment, s, is processed into a dimension reduced property vector $\mathbf{P}$, from which a prediction, $\hat{D}(\tau_s)$, of the coding distortion, $D(\tau_s)$ is to be found. For simplicity, we below drop segment and coding template indices. The random variable representing the incurred distortion will be denoted $\mathcal{D}$, and the corresponding outcomes will be denoted δ.

The selection of a set of properties, $\mathbf{p}$, from the input segment, s, is of great importance for the performance of the proposed framework. The selected set of properties should be a representative for the incurred distortion in the current segment for the given coder. In more theoretical terms, the random input segment, s, is processed into two random variables, the distortion variable, $\mathcal{D}$, with outcomes δ, and the property vector, $\mathbf{P}$. The basic task for the property extractor, $f(\cdot)$, is to extract properties, $\mathbf{P}$, that contain sufficient information about $\mathcal{D}$ for a required predictor accuracy. The amount of information that $\mathbf{P}$ contains about $\mathcal{D}$, or the goodness of a given property vector, can be measured by the mutual information $I(\mathcal{D}; \mathbf{P})$. In this work we have chosen to rely on standard audio properties. Our choice of property vector is further discussed in Section 4.

The aim of the predictor, $g(\cdot)$, is to find a prediction, $\hat{\delta}$, of the incurred distortion, δ, based on an observation of the property vector, $\mathbf{P} = \mathbf{p}$. Utilizing a pre-trained GMM for the joint distortion property pdf, $f^{(\mathcal{M})}_{\mathcal{D},\mathbf{P}}(\delta, \mathbf{p})$, we approximate the MMSE at each coding instant as

$$\hat{\delta} = g(\mathbf{p}) = \int \delta f^{(\mathcal{M})}_{\mathcal{D}|\mathbf{P}}(\delta|\mathbf{P} = \mathbf{p}) d\delta, \tag{7}$$

where $f^{(\mathcal{M})}_{\mathcal{D}|\mathbf{P}}(\delta|\mathbf{P} = \mathbf{p})$ is the conditional model pdf, which can be shown to be a mixture of Gaussian densities, and is easily derived from the joint model pdf, $f^{(\mathcal{M})}_{\mathcal{D},\mathbf{P}}(\delta, \mathbf{p})$.

In practice, this predictor calculates a weighted sum of conditional means,

$$\hat{\delta} = \sum_{i=1}^{M} \rho_i' \mathbf{m}_{i,\mathcal{D}|\mathbf{P}=\mathbf{p}}, \tag{8}$$

where M is the number of mixture components, and $\{\rho_i'\}$ and $\{\mathbf{m}_{i,\mathcal{D}|\mathbf{P}=\mathbf{p}}\}$ represent the weights and the means of the conditional model pdf, $f^{(\mathcal{M})}_{\mathcal{D}|\mathbf{P}}(\delta|\mathbf{P}=\mathbf{p})$, respectively.

3.2 Performance

The employed prediction scheme is designed to minimize the variance of the prediction error, $Z = \delta - \hat{\delta}$. Assuming an unbiased predictor, the variance of the prediction error can be expressed as

$$\sigma_Z^2 = \mathrm{E}\left[(Z)^2\right] = \mathrm{E}\left[(\delta - \hat{\delta})^2\right]. \tag{9}$$

The minimum mean square error estimator (MMSE) for this task, i.e., the one minimizing σ_Z^2, is the conditional mean estimator,

$$\hat{\delta}_{\mathrm{mmse}} = \mathrm{E}[\mathcal{D}|\mathbf{P}=\mathbf{p}] = \int \delta f_{\mathcal{D}|\mathbf{P}}(\delta|\mathbf{P}=\mathbf{p}) d\delta. \tag{10}$$

The employed predictor is an approximation of the MMSE estimator, and the predictor output (8) will approach the true conditional (10), as the model pdf approaches the true pdf.

As discussed above, the performance of the predictor is dependent on the chosen property vector. In [5] the relation between the property goodness, $I(\mathcal{D};\mathbf{P})$, and the overall prediction error, σ_Z^2 was studied. It was shown that for a given property vector, $\mathbf{P}$, the overall prediction error, σ_Z^2, can be bounded as

$$\sigma_{\mathcal{D}}^2 \geq \sigma_Z^2 \geq \frac{1}{2\pi e} 2^{2(h(\mathcal{D}) - I(\mathcal{D};\mathbf{P}))}, \tag{11}$$

where $\sigma_{\mathcal{D}}^2$ is the variance of the distortion variable to be predicted, $h(\mathcal{D})$ is the differential entropy of the distortion random variable $\mathcal{D}$, and $I(\mathcal{D};\mathbf{P})$ is the mutual information between $\mathcal{D}$ and $\mathbf{P}$.

4 Experimental Setup

Here, we present the experimental framework, separated into the source coding system (sinusoidal coder, R-D optimization , distortion criterion), and the distortion predictor (GMM, property vector, audio database).

4.1 Source Coding System

We employ a sinusoidal coder based on a simplified version of psychoacoustic matching pursuit (PAMP) [7]. Using a PAMP based coder, the distortion (12) will decrease in a monotone way as a function of the number of iterations (sinusoids). The analysis/-synthesis is performed for segments of length 35 ms, sampled at 48 kHz. The coder employs a Hanning window and has a 50 % segment overlap. Phases are quantized uniformly using 5 bits per component, whereas amplitudes and frequencies are quantized in the logarithmic domain. Using entropy coding and differential encoding, we obtain perceptually transparent quantization at an average rate of approximately 16 bits/sinusoid.

R-D optimization, c.f. Section 2, is here employed to distribute sinusoids (bit-allocation) over optimization viewports, $\mathcal{S}$, of length 1 s, matching a target rate of 25 kbits/s[2]. For each segment the algorithm allocates a number of sinusoids in the range $0-85$. The optimization is performed using the sinusoidal modeling distortion as input, using a cost of 16 bits/sinusoid.

We employ a distortion criterion, $\delta(\cdot)$, based on the model in [6]. For a particular segment the distortion can be written as

$$\delta(e(n)) = \int_{-\pi}^{\pi} A(\omega)|\mathcal{F}\left\{w(n)e(n)\right\}|^2 d\omega, \tag{12}$$

where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform, $A(\omega) \in \{x \in \mathbb{R} | x > 0\}$ is a perceptual weighting function and $w(n)$ is the analysis window and $e(n) = \tilde{s}(n) - s(n)$ is the modeling error. The quantization distortion is disregarded in the optimization as the distortion criterion may be overly sensitive to frequency quantization.

4.2 Distortion Predictor

The key component of the predictor described in Section 3 is a GMM for the joint property-distortion pdf, $f^{(\mathcal{M})}_{\mathcal{D},\mathbf{P}}(\delta, \mathbf{p})$, which is to be trained off-line. All GMM's employ 16 mixtures, and the training was conducted using the expectation maximization-algorithm (EM). For GMM training purposed we have extracted a training set, consisting of 180.000 joint property-distortion vectors from the SQAM database (up-sampled to 48 kHz). All test excerpts are disjoint from the training set.

We have chosen to work with a 4-dimensional property vector consisting of: 1) The loudness, which is calculate as the log of the average energy of the segment, 2) The spectral centroid, which is calculated as mean of the spectrum with respect to frequency, 3) Spectral bandwidth, which is calculated as the second moment of the spectrum with respect to frequency, 4) Spectral flatness, which is calculated as the ratio of the geometric mean and the arithmetic of the power spectrum. We do not claim to

[2] In this context coding templates, referred to in Section 2, correspond to a sinusoidal coder using different number of sinusoids.

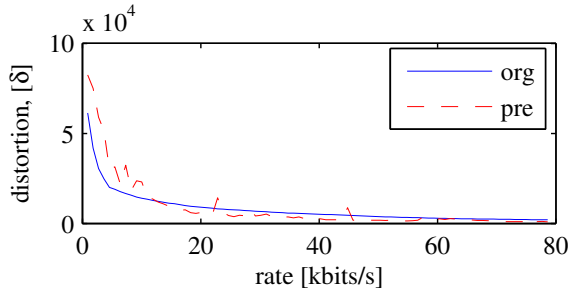

Figure 3: Original and predicted R-D curves for one segment (35 ms) in the excerpt "glockenspiel".

excerpt	$\mathrm{E}[\delta_{\mathrm{org}}]$	$\Delta\mathrm{E}[\delta_{\mathrm{pre}}]$	$\Delta\mathrm{E}[\delta_{\mathrm{uni}}]$
glockenspiel	$6.55 \cdot 10^2$	50 %	103 %
german speech	$2.42 \cdot 10^4$	3.8 %	9.9 %
castanets	$2.68 \cdot 10^4$	2.7 %	7.6 %
harpsichord	$9.63 \cdot 10^3$	7.6 %	18 %
jazz	$2.79 \cdot 10^4$	3.2 %	4.2 %

Table 1: Average segment distortion, $\mathrm{E}[\delta_{\mathrm{org}}]$, for various excerpts. $\Delta\mathrm{E}[\delta_{\mathrm{sys}}] = \frac{\mathrm{E}[\delta_{\mathrm{sys}}]-\mathrm{E}[\delta_{\mathrm{org}}]}{\mathrm{E}[\delta_{\mathrm{org}}]}$ represents the increase in average distortion compared to an R-D optimized system based on original R-D curves. Here shown for a system using predicted R-D curves(pre), and for a system using uniform bit allocation over the segments (uni).

have chosen the best property for the task at hand, rather we have chosen to rely on simple (low-complexity) standard audio properties, used in audio classification [8].

5 Experimental Results

We have tested the proposed open loop R-D optimization, for the purpose of R-D optimized bit-allocation (distribution of sinusoids) over optimization viewports, $\mathcal{S}$, c.f. Section 2. In the experiments we have exchanged original R-D pairs, c.f. equation (6), with predicted R-D pairs. For our particular setup, this means that we have exchanged 86 original R-D pairs, below referred to as a R-D curve, with predicted ones for each segment, as visualized in Figure 3. Predicted distortion values are only used in the optimization, meaning that presented distortion values are based on original R-D curves. For comparison purposes we have included the performance of a coder with a uniform sinusoidal distribution, i.e. the same number of sinusoids per segment.

In Table 1 we compare the performance of the systems, by averaging the distortion in equation (12) over a number of different excerpts. The results show that the proposed

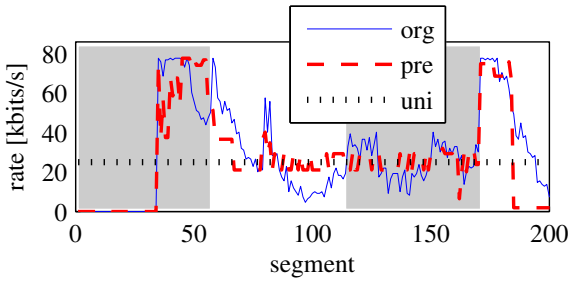

Figure 4: Bit allocation for the first 200 segments of the excerpt "Glockenspiel". The solid line represents an R-D optimized system based on original R-D curves (org), the dashed line represents an R-D optimized system based on predicted R-D curves (pre), and the dotted line represent a system with uniform bit allocation (uni). The periodically changing shaded and white fields represents optimization viewports.

system outperforms a uniform sinusoidal distribution for all the excerpts. Naturally, there is a loss compared to the reference system. The gains of optimized systems compared to a system using a fixed number of sinusoids vary. The achievable gain of R-D optimized coding is large for "glockenspiel". The non-stationary character of the signal, results in an R-D optimized bit distribution which is far from uniform, c.f. Figure 4. The result is a far too high distortion at onsets for the uniform case, c.f. Figure 5. For the "jazz" excerpt the R-D optimized distribution of sinusoids is not far from uniform, and thus a uniform distribution can compete with the RD optimized, c.f. Table 1. It should be mentioned that the poor performance for the proposed system on the "glockenspiel" excerpt, a 50 % loss, can be traced back to the R-D optimization procedure. Due to the non-convexity of predicted RD-curves, c.f. Figure 3, the optimization fails in selecting the correct operating point. By simple post processing of predicted R-D curves, smoothing and forcing convexity, the loss can easily be reduced to around 20 %.

An alternative application is up-front coder selection for each optimization viewport, $\mathcal{S}$, i.e. selection of the coder that minimizes the distortion for the current set of segments, $\mathcal{S}$. For this purpose viewport R-D curves are useful. Viewport R-D curves are achieved by sweeping over $\lambda^\star$, c.f. equation (4). In Figure 6 R-D curves for the first viewport in the "glockenspiel" excerpt are shown. The solid line represents the viewport R-D curve based on original distortion values, the dashed line represents the predicted viewport R-D curve and the dotted line represent the R-D curve for a sinusoidal coder employing a fixed number of sinusoids per segment. Comparing the solid and the dotted curve indicate that we should select the R-D optimized system instead of the fixed system for all rates on this viewport. We can also note that the choice would be the same if we based our decision on the predicted curve, the dashed line, instead of the original. This is obviously a dummy selection, as an optimized system always

outperforms a fixed system, but if the dotted curve would have represented for example a waveform coder, such a selection can be of interest. Note that the dashed line represents a prediction of the performance of the R-D optimized system (solid line), and it can therefore indicate a performance better than the performance of the actual system[3].

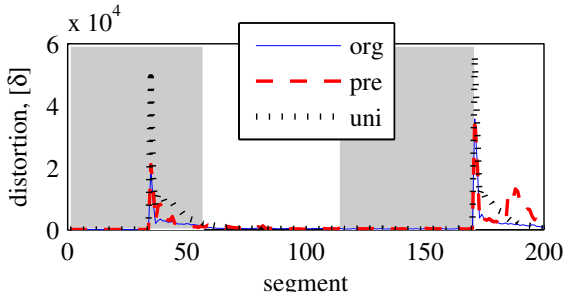

Figure 5: Distortion distribution for 200 segments of the excerpt "Glockenspiel". The solid line represents an R-D optimized system based on original R-D curves (org), the dashed line represents an R-D optimized system based on predicted R-D curves (pre), and the dotted line represent a system with uniform bit allocation (uni). The periodically changing shaded and white fields represents optimization viewports.

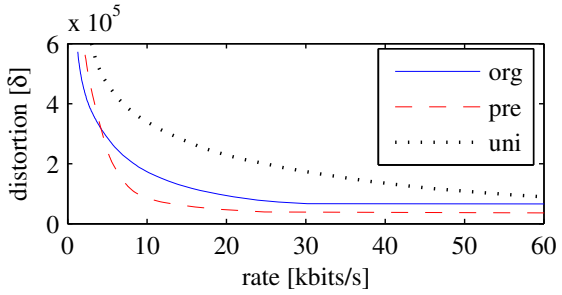

Figure 6: Viewport R-D curves for the first optimization viewport (1 s) in the excerpt "glockenspiel". The solid line represents an R-D optimized system based on original R-D curves (org), the dashed line represents an R-D optimized system based on predicted R-D curves (pre), and the dotted line represent a system with uniform bit allocation (uni).

[3]Here all figures are based on predicted distortion values, as opposed to above, where predicted distortion values only are used for the optimization, and the results are based on original distortion values.

6 Discussion

In this paper we have studied complexity reduced R-D optimized coding, where R-D curves are exchanged for predicted ones. The proposed framework was applied in a sinusoidal coding context, for the purpose of distributing sinusoids over sets of audio segments. The results show that the proposed framework works, in the sense that the performance is improved compared to a system with a uniform sinusoidal distribution. It should be noted that we lose compared to an R-D optimized system based on the true R-D curves. This loss can be decreased if our rather raw system is further optimized, meaning a better choice of property vector, and a set of training data better matching the expected audio input.

References

[1] M. G. Christensen and S. van de Par, "Rate-distortion efficient amplitude modulated sinusoidal audio coding," in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2280–2284.

[2] R. Vafin and W. B. Kleijn, "Towards optimal quantization in multistage audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2004, pp. 205–208.

[3] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[4] P. Prandoni, "Optimal Segmentation Techniques for Piecewise Stationary Signals," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne, 1999.

[5] F. Nordén, S. V. Andersen, S. H. Jensen, and W. B. Kleijn, "Property vector based distortion estimation," in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004, pp. 2275–2279.

[6] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2001, pp. 1805 – 1808.

[7] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[8] E. Wold *et al.*, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

Paper H

Low Complexity Rate-Distortion Optimized Time-Segmentation for Audio Coding

Christoffer A. Rødbro, Mads Græsbøll Christensen, Fredrik Nordén, and Søren Holdt Jensen

The paper has been published in
Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 231–234, 2005.

Abstract

In this paper, we investigate a reduced complexity approach to rate-distortion optimized time-segmentation in audio coding. Instead of the conventional closed-loop approach for determining the coding distortions, they are estimated from a set of features extracted from the audio signal. Care is taken to ensure that properties such as convex and non-increasing rate-distortion curves carry over from the training data to the estimated rate-distortion pairs. With computational complexity reductions of a factor close to 10, perceptual listening tests reveal a slight reduction of the signal quality, while maintaining a large improvement over fixed segmentation.

1 Introduction

Adaptive time-segmentation has been shown to be an efficient method for improving the rate-distortion trade-off in speech- and audio coding [1–4]. These methods usually employ an analysis-by-synthesis procedure in which full encoding-decoding operations are required for each and every candidate segment, including those not actually used in the final signal representation. This is necessary in order to determine the distortion and rate if the segment is used in the signal representation. The distortions are obtained by explicitly comparing the encoded-decoded segments to the corresponding original segments, and the optimal segmentation is then found as the one minimizing the total distortion, usually subject to a rate constraint. If the encoding-decoding processes are computationally extensive, as is the case e.g. in the psychoacoustic matching pursuits (PMP) schemes of [5, 6], this may lead to impractical execution times, even for off-line applications such as audio compression. However, [7] proposed a strategy for estimating, at low complexity, the distortion arising from coding a signal segment. In [8], this approach was used to predict the optimal distribution of sinusoidal components in a fixed segmentation PMP coder. In this work we shall use a slightly modified approach to estimate the optimal time-segmentation in the same coder.

The rest of this paper is structured as follows: first, rate-distortion optimized time-segmentation is reviewed in Section 2. Next, Sections 3 and 4 describe how to incorporate the distortion estimation approach of [7] into such a scheme. Objective as well as subjective results are given in Section 5 before Section 6 concludes on the work.

2 Rate-Distortion Optimized Time-Segmentation

The rate-distortion optimized time-segmentation algorithm of [1] is based on the constrained optimization problem:

$$\begin{aligned} \text{minimize}: \quad & D(\tau, \mathbf{c}(\tau)) \\ \text{s.t.}: \quad & R(\tau, \mathbf{c}(\tau)) \leq R_C. \end{aligned} \tag{1}$$

Here, $\tau = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{\sigma(\tau)}\}$ denotes the time-segmentation consisting of $\sigma(\tau)$ variable length segments $\mathbf{s}_i$, each having a length equal to an integer number of grids (e.g. 5 ms). The vector $\mathbf{c}(\tau) = \{c_1(\tau), c_2(\tau), \ldots, c_{\sigma(\tau)}(\tau)\}$ denotes the coding templates, (i.e. different ways of encoding each segment in a segmentation τ). R_C is the target bit budget, whereas R is the total number of bits used and D is the total distortion, the latter two found by summation over the segments:

$$R(\tau, \mathbf{c}(\tau)) = \sum_{i=1}^{\sigma(\tau)} r(c_i(\tau)) \quad \text{and} \quad D(\tau, \mathbf{c}(\tau)) = \sum_{i=1}^{\sigma(\tau)} d(c_i(\tau)).$$

Here, $r(c_i(\tau))$ is the number of bits used for encoding segment $\mathbf{s}_i$ using template $c_i(\tau)$ and $d(c_i(\tau))$ is some measure of the distortion between the original segment and the one encoded using template $c_i(\tau)$. Usually, the constrained optimization problem (1) is solved by recasting it as an unconstrained problem with cost-function:

$$J(\tau, \mathbf{c}(\tau)) = D(\tau, \mathbf{c}(\tau)) + \lambda R(\tau, \mathbf{c}(\tau)). \tag{2}$$

Now, by setting λ to some value (say λ_x) and minimizing J over $\{\tau, \mathbf{c}(\tau)\}$ we will obtain a pair (D_x, R_x) optimal for λ_x. Thus, λ can be iterated and J minimized in each step, until a rate $R \lessapprox R_C$ is obtained. In each iteration, the minimization of J is a two-step procedure: first, the coding templates $c_i^\star(\tau)$ optimal for λ_x are found for each segment:

$$\forall i, \tau : c_i^\star(\tau) = \arg\min_{c_i(\tau)} \left\{ d(c_i(\tau)) + \lambda_x r(c_i(\tau)) \right\}. \tag{3}$$

By denoting $j_i^\star(\tau) = d(c_i^\star(\tau)) + \lambda r(c_i^\star(\tau))$, the optimal segmentation $\tau^\star$ is the one minimizing the sum over $j_i^\star(\tau)$:

$$\tau^\star = \arg\min_{\tau} \sum_{i=1}^{\sigma(\tau)} j_i^\star(\tau). \tag{4}$$

This minimization is carried out at reasonable complexity using a dynamic programming technique, see [1] for details.

The computational problems of the procedure described above appears in (3): in order to find the optimal coding templates, we need the distortion if the coding template is used in the signal representation. Thus, we must encode all segments with all coding templates, even for the segments and coding templates not used in the final representation. If we denote the number of grids in the signal by G and all segment lengths from 1 to G grids are allowed, the total number of possible segments in the signal equals $K = \frac{G^2+G}{2}$. Alternatively, if the maximum segment length is limited at L with $G \gg L$, $K \approx GL$. This is significantly larger than the number of segments actually used in the signal representation, $\sigma(\tau) \leq G$.

The number and nature of the coding templates depends directly on the type of coder(s) employed. In the rest of this paper, we shall focus on (psychoacoustic) MP-based coders, e.g. [5, 6, 8]. In such a coder, the signal segments are iteratively decomposed into a weighted sum of basis functions. In each iteration, a basis function is chosen from an over-complete dictionary as the one minimizing a perceptual error norm (a distortion). In that the representation of each basis function requires a certain amount of bits, varying the number of components results in different coding templates with corresponding $\{r(c_i(\tau)), d(c_i(\tau))\}$ pairs. Due to the MP approach, these pairs will lie on a non-increasing (and sometimes also convex) hull. Unfortunately, the PMP algorithm is computationally extensive, primarily because accurate modeling of certain signal segments such as transients requires quite a large number of iterations, (e.g. [8] applied 0-85 sinusoids in each segment). Even with the efficient implementation of [6] requiring 3 FFTs per iteration, this results in up to 255 high-order FFTs for each of the K frames; clearly, means for reducing this complexity is called for. One way of doing so is based on the observation that running the MP on all K possible segments is wasteful, because only $\sigma(\tau)$ of them are used in the final segmentation. This motivates estimating the distortions $d(c_i(\tau))$ used in (3) instead of calculating them explicitly.

3 Distortion Estimation

The principle of [7] is to estimate the coding distortion from a vector of features extracted from each candidate audio segment; the computational complexity required to determine these features should be low, or little complexity reduction is gained. Section 4 will account for the features explicitly used, but they should be descriptive of the signal, such as spectral information, periodicity, stationarity, power, etc. The P features are stacked in a vector $\mathbf{p}_i$, i denoting the candidate segment index. Now, the distortions arising if assigning $1, 2, ..., C$ components for representing the segment are added to this vector[1]:

$$\mathbf{o}_i = \begin{bmatrix} d_i^{(1)} & d_i^{(2)} & \cdots & d_i^{(C)} & \mathbf{p}_i^T \end{bmatrix}^T = \begin{bmatrix} \mathbf{d}_i \\ \mathbf{p}_i \end{bmatrix} \in \mathbb{R}^{C+P} \tag{5}$$

Now, from a set of training data (we used a subset of the SQAM database [9]), a pdf in the form of a multivariate Gaussian mixture is estimated:

$$\mathbf{o}_i \sim \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{6}$$

where M is the number of mixture components, w_m denotes the mixture weights ($\sum_{m=1}^{M} w_m = 1$), and $\boldsymbol{\mu}_m$, $\boldsymbol{\Sigma}_m$ are the Gaussian mean vectors and covariance ma-

[1] Actually, the vector is built from normalized distortions, $d_i^{(c)}/\|\mathbf{s}_i\|_2^2$, with the estimates being rescaled accordingly. This reduces the dynamic range of the distortions and thus eases the statistical modeling to be described in the following.

trices, respectively. w_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are found using the expectation maximization algorithm [10], with a model being build for each possible segment length. In the following, we shall drop the subscript i leaving the frame index implicit.

At this point, we have obtained a pdf in the form of a Gaussian Mixture Model (GMM) describing the features and the distortion arising from coding jointly. The task at hand is: given the features extracted from a segment and the GMM, estimate the vector of distortions. It can be shown that the conditional MMSE estimator is of the form:

$$\hat{\mathbf{d}} = E[\mathbf{d}|\mathbf{p}] = \sum_{m=1}^{M} \tilde{w}_m \tilde{\boldsymbol{\mu}}_m, \tag{7}$$

where $0 \leq \tilde{w}_m \leq 1$ depends on $\mathbf{p}$ (see [7] for details), and

$$\tilde{\boldsymbol{\mu}}_m = \boldsymbol{\mu}_{m,d} + \boldsymbol{\Sigma}_{m,dp} \left(\boldsymbol{\Sigma}_{m,pp}\right)^{-1} \left(\mathbf{p} - \boldsymbol{\mu}_{m,p}\right), \tag{8}$$

with $\boldsymbol{\mu}_{m,d} \in \mathbb{R}^C$ and $\boldsymbol{\mu}_{m,p} \in \mathbb{R}^P$ being sub-vectors of $\boldsymbol{\mu}_m$,

$$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_{m,d} \\ \boldsymbol{\mu}_{m,p} \end{bmatrix}, \tag{9}$$

whereas $\boldsymbol{\Sigma}_{m,dp} \in \mathbb{R}^{C \times P}$ and $\boldsymbol{\Sigma}_{m,pp} \in \mathbb{R}^{P \times P}$ are sub-matrices of $\boldsymbol{\Sigma}_m$,

$$\boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{\Sigma}_{m,dd} & \boldsymbol{\Sigma}_{m,dp} \\ \boldsymbol{\Sigma}_{m,pd} & \boldsymbol{\Sigma}_{m,pp} \end{bmatrix}. \tag{10}$$

In some cases, the approach reviewed above leads to a problem reported in [8], in that there is no guarantee that the estimated distortion vector $\hat{\mathbf{d}}$ will be a non-increasing sequence. This leads to certain problems, for example, the algorithm does not recognize that adding sinusoidal components never leads to increased distortion. However, this problem is easily circumvented by confining the covariance matrices to a diagonal structure, implying that $\boldsymbol{\Sigma}_{m,dp} = \mathbf{0}$ so that (8) reduces to:

$$\tilde{\boldsymbol{\mu}}_m = \boldsymbol{\mu}_{m,d}. \tag{11}$$

Now, the estimator in (7) is a positively weighted sum of the GMM mean sub-vectors $\boldsymbol{\mu}_{m,d}$ and thus non-increasing if the individual $\boldsymbol{\mu}_{m,d}$ are. This is indeed true because in the EM-algorithm, the $\boldsymbol{\mu}_m$ updates are positively weighted sums of the training vectors [10]. Thus, because the distortion vectors $\mathbf{d}_i$ extracted for training are non-increasing, so are $\boldsymbol{\mu}_{m,d}$, and in turn $\hat{\mathbf{d}}$. Also, note that convexity carries over in the same way, which is a coveted property because it prevents ambiguity in the minimization (3).

Also, it should be noted that constraining the covariance matrices to be diagonal has the beneficial side effect of significantly reducing the computational complexity associated with finding $\tilde{w}_m$ and $\tilde{\boldsymbol{\mu}}_m$. Specifically, the main complexity in calculating $\tilde{w}_m$ stems from evaluating M Gaussians in the GMM, which has complexity $\mathcal{O}(MP^2)$

for full covariance matrices, but only $\mathcal{O}(MP)$ for diagonals. Also for full covariance matrices, determining $\tilde{\boldsymbol{\mu}}_m$ for all m using (8) has complexity $\mathcal{O}(MPC)$, whereas the diagonal case of (11) is cost free. On the other hand, a somewhat larger number of mixtures M will be necessary to obtain a precise model.

4 The Feature Vector

A problem not addressed in the preceding work is how to select which features to include in the vector $\mathbf{p}$. For this, a "deflation" strategy is employed, the idea being to start out with a large number of parameters and then sequentially remove one at a time until the estimation performance begins to degrade on a test set. Such a large number of parameters requires many degrees of freedom in the model, and we therefore used $M = 320$ mixture components. The initial length $P = 22$ parameter vector contained the parameters listed in Table 1.

Number	Description	Used
1.	Signal power	No
2.	Number of zero-crossings	No
3.	Loudness (log-power)	Yes
4.	A spectral flatness measure	Yes
5.	A spectral centroid measure	Yes
6.	A spectral bandwidth measure	Yes
7.	An LPC flatness measure	Yes
8.	A periodicity measure	No
9.-20.	12 mel-cepstrum coefficients	No
21.	A power stationarity measure	Yes
22.	A spectral stationarity measure	Yes

Table 1: The features included in the initial feature vector $\mathbf{p}$. The used column indicates whether the feature was used in the listening test.

An example illustrating the behavior of the deflation strategy is shown in Figure 1. The left-hand plot seems to indicate that no feature is much more important than any other; the estimated distortion MSEs obtained are quite similar. However, since the case where the signal power is removed gives a slightly better overall performance, this parameter is eliminated. Then, in the next iteration, parameter number 3 (log-power) becomes very important, since the information contained in this parameters is no longer redundant with the rest. Also, this plot indicates that the next parameter to be removed from the model should be number 8, the periodicity measure. Using this approach, the features sequentially removed from the parameter vector were the mel-cepstrum coefficients, the signal power, the periodicity measure, and the number of

zero-crossings, resulting in a final parameter vector length of $P = 7$. It should be noted that different parameters (and coders) could be applied for different segment lengths; doing so, however, is beyond the scope of this paper.

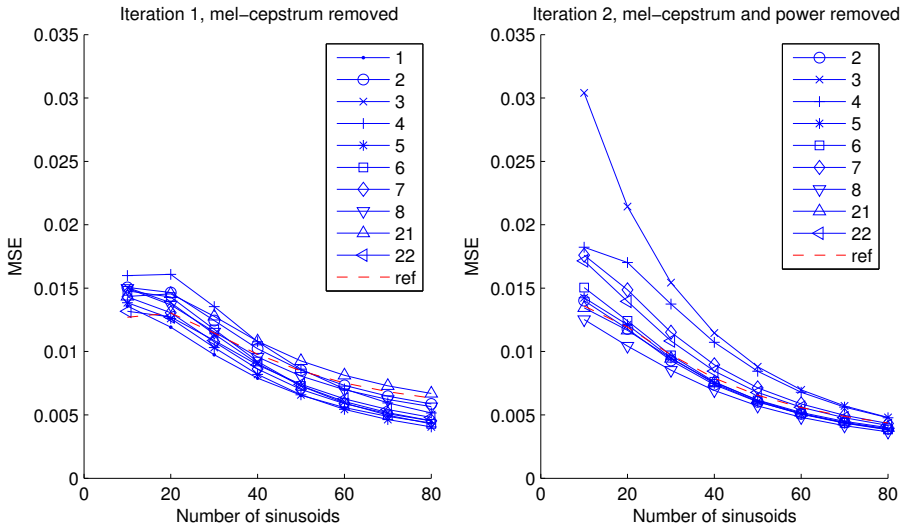

Figure 1: Example illustrating the deflation strategy for a segment length of 30 ms. The legend number refers to which feature has been removed, corresponding to the numbers in Table 1. "ref" represents the MSE when none of the parameters are removed.

5 Experiments

In the following, experimental results will be presented with 4 different segment lengths being allowed in the segmentation: 10 ms, 20 ms, 30 ms and 40 ms (including 5 ms overlap). For fixed segmentation, a window update rate of 15 ms was used, corresponding to the 20 ms window in adaptive segmentation. Through informal listening, these windows were found appropriate for the 30 kbps target bit rate used. For further details, see [8].

An example of the optimal and the estimated segmentations is shown in Figure 2 for a section of the SQAM "claves" signal. We see that the estimation captures the onset, whereas the segmentation deviates in the more stationary signal areas. This is a typical behavior that seems sensible, in that adaptive segmentation has its greatest impact in non-stationary signal areas.

To assess the perceptual degradations (if any) induced by the proposed segmentation approach as compared to optimal segmentation a MUSHRA [11] listening test was

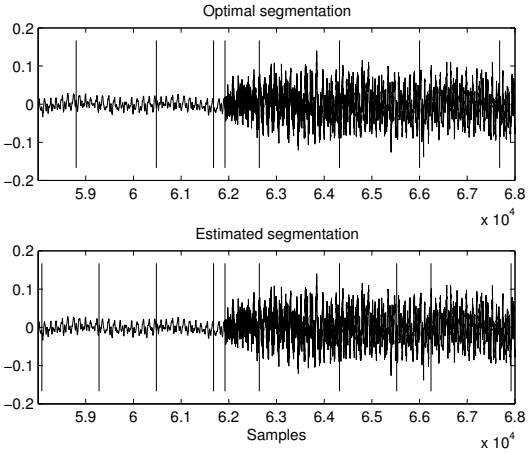

Figure 2: Comparison of the optimal segmentation and that yielded by the proposed method. The vertical lines represent the segmentation.

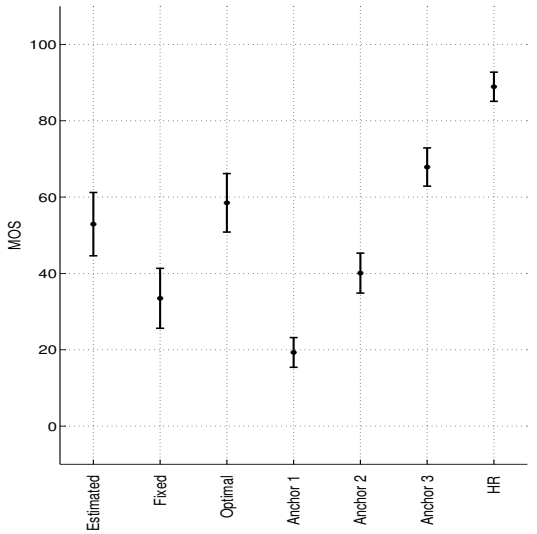

Figure 3: Average scores of MUSHRA listening test. The vertical lines indicate the 95% confidence intervals.

carried out. The test set consisted of 6 different audio samples (3 single instrument, 1 solo, 1 orchestra, and 2 pop), none of which were included in the training. The samples

were presented for the proposed method based on estimated distortions ("estimated"), for fixed segmentation ("fixed"), and for rate-distortion optimal segmentation ("optimal"). Moreover, signals low-pass filtered at 3.5 kHz, 7 kHz and 10 kHz were included as anchors 1 to 3, whereas the original was included as a hidden reference (HR). Averaged results for 8 listeners are shown in Figure 3. Typically, the ratio between the total number of segments, K, and the number of segments actually used, $\sigma(\tau)$, lay between 10 and 15.

6 Conclusion

The scores in Figure 3 indicate that the perceptual quality is slightly degraded for the distortion estimation based approach as compared to optimal segmentation. However, there is still a significant quality gain over fixed segmentation. These results should be compared to computational complexity of the methods. While the optimal segmentation approach requires K executions of the PMP, the distortion estimation based approach requires only $\sigma(\tau)$, with the ratio $\frac{K}{\sigma(\tau)} > 10$. On top of this, the distortion estimation approach requires a feature vector extraction as well as the GMM-based estimation procedure described in Section 3 for each of the K segments. However, the complexity of these steps is low compared to the +200 FFTs required by the PMP, so realistically the complexity reduction is in the neighborhood of 10. This is supported by the observed MATLAB execution times.

References

[1] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech and Audio Processing*, pp. 646–655, 8(6) 2000.

[2] P. Prandoni, M. M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 2029–2032.

[3] C. A. Rødbro, J. Jensen, and R. Heusdens, "Adaptive time-segmentation for speech coding with limited delay," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2004, pp. 465–468.

[4] M. G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(4), July 2006.

[5] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1809–1812.

[6] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling of audio and speech using psychoacoustic-adaptive matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 3281–3284.

[7] F. Nordén, S. V. Andersen, S. H. Jensen, and W. B. Kleijn, "Property vector based distortion estimation," in *Rec. Thirty-Eighth Asilomar Conf. Signals, Systems, and Computers*, 2004.

[8] F. Norden, M. G. Christensen, and S. H. Jensen, "Open loop rate-distortion optimized audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2005, pp. 161–164.

[9] European Broadcasting Union, *Sound Quality Assessment Material Recordings for Subjective Tests*. EBU, Apr. 1988, Tech. 3253.

[10] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001, ch. 4, pp. 172–175.

[11] *ITU-R BS.1534*, ITU Method for subjective assessment of intermediate quality level of coding system, 2001.

Paper I

Compressed Domain Packet Loss Concealment of Sinusoidally Coded Speech

Christoffer A. Rødbro, Mads Græsbøll Christensen,
Søren Vang Andersen, and Søren Holdt Jensen

The paper has been published in
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. I, pp. 104–107, 2003.

Abstract

In this paper we consider the problem of packet loss concealment for Voice over IP (VoIP). The speech signal is compressed at the transmitter using a sinusoidal coding scheme working at 8 kbit/s. At the receiver, packet loss concealment is carried out working directly on the quantized sinusoidal parameters, based on time-scaling of the packets surrounding the missing ones. Subjective listening tests show promising results indicating the potential of sinusoidal speech coding for VoIP.

1 Introduction

In packet-switched communication systems, such as the Internet, packets may be delayed or even lost during transmission. This is not critical in most applications since the receiving end can request retransmission of the packet in question. However, in a real-time constrained application such as VoIP, retransmission is not feasible since this would introduce a considerable delay prohibiting proper two-way conversation. Thus lost and delayed packets must be compensated for at the receiving end. This is usually attempted by storing a number of recently arrived packets in a jitter buffer before playout. If the packet delay is lower than the time extension of the jitter buffer it can be used to compensate for packet delay variations (jitter). However, packets delayed more than the length of the jitter buffer are considered lost and have to be replaced.

The simplest approaches in case of packet loss are silence or noise substitution but these methods have a highly negative impact on perceived speech quality. Better approaches rely on waveform substitution from neighboring frames, see e.g. [1]. More recently, missing frames were estimated through a combination of LPC analysis and interpolation/extrapolation of the residual signal using sinusoidal modeling [2], [3]. Instead of estimating the missing packet, another approach is to stretch the packets preceding the missing one in order to allow more time for delayed packets to arrive [4], [5]. In a VoIP application the speech signal would normally be compressed to achieve a lower bit rate. An important design criterion for such speech coding schemes is robustness towards packet losses, see e.g. [6]. Moreover, the data made available by the speech coder at the receiver should be sufficient to facilitate packet loss concealment.

In this paper we utilize a speech coding algorithm based on sinusoidal modeling which is described in Section 2. In Section 3 we then propose a packet loss concealment algorithm based on time-scale modification which works directly on the sinusoidal parameters. The sinusoidal coding scheme is a modified version of that presented in [7] whereas the packet loss concealment is based on [8]. Experimental results are presented and discussed in Section 4 before Section 5 concludes on the work.

2 Sinusoidal Coder

Speech coding for use in packet switched networks should be designed for robustness towards packet losses. One way of achieving this is to ensure that decoding of frames can be performed independently. Also, it is desirable to design the coder in such a way that it is possible to perform packet loss concealment in the compressed domain. These properties can easily be incorporated into a sinusoidal coder. We have developed a fixed bit-rate sinusoidal coder operating at 8 kbit/s suitable for packet switched networks as a reference system for testing the packet loss concealment method proposed. This is done to ensure that the method can operate under realistic conditions with quantized parameters. The coder of [7] has been modified to fit the requirements of packet switched networks. It is based on a harmonic sinusoidal model, where the speech segment is represented as a finite sum of harmonically related sinusoids:

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_0 l n + \phi_l) \tag{1}$$

Here ω_0 is the fundamental frequency and L is the number of components in the segment, and A_l and ϕ_l are the amplitude and phase of the l'th harmonic respectively. After segmentation the parameters of this model are estimated. Here, the speech is split into segments of 20 ms with 50% overlap. First, the pitch is estimated using the correlation based method proposed in [9]. The problem of finding the optimum amplitudes and phases then turns into a linear least-squares problem that is solved using weighted least squares (WLS), see e.g. [8] for details. Although the harmonic sinusoidal model is only physiologically founded for voiced speech, it is well-known that it can be used for modeling of noise-like signals [10] such as unvoiced speech, provided that the frequency spacing is sufficiently small. A frequency spacing of 100 Hz for unvoiced speech has been found to form a reasonable tradeoff between model performance and the number of parameters. The cumulative mean normalized difference function in [9] is used for voiced/unvoiced decision and to estimate a voicing dependent cut-off frequency, ω_c. The amplitudes are represented using a 10th order discrete all-pole model (DAP) [11]. In this model the spectral envelope is optimized to match only at the discrete harmonic frequencies rather than the continuous spectrum. It is then represented using line spectral frequencies and finally "transparently" coded using perceptually weighted split vector quantization with a 24 bit codebook as described in [12]. The fundamental frequency and the gain are quantized in the log-domain using 7 and 5 bits respectively. The phases can be represented efficiently by exploiting the near-linear relationship between the phases of the harmonics of voiced speech. This has been done by fitting a line to the unwrapped phases and the parameters of the line are encoded using a total of 7 bits. As the phases are only approximately linear and only in perfectly voiced regions, there are non-zero phase residuals or errors. These are then quantized using a scalar uniform quantizer in the range $]-\pi, \pi]$. Bits are allocated in accordance with the power distribution (the quantized DAP envelope) such that

the phases of the largest components receive more bits than smaller ones. In unvoiced regions the phases are simply quantized directly. The reason for using bits for phase quantization in unvoiced segments is that it provides better modeling as waveform approximating capabilities are achieved. This is important in e.g. the burst of a plosive, where the phases are not stochastic. Also, it has been found to generally improve the perceived quality as well as improving robustness due to the waveform approximating property. In Table 1 the bit allocation per frame of the coder for operation at 8 kbit/s is shown. In the decoding process phase randomization inversely proportional to the

Parameter	Voiced	Unvoiced
V/UV	1	1
Pitch	7	0
Linear Phase Coefs	7	0
Cut-off Frequency	2	0
Phase Residuals	34	50
LP Gain	5	5
LSF VQ Index	24	24
Total	80	80

Table 1: Fixed rate bit allocation (per frame).

number of bits allocated for a given component should be applied with different ranges depending on the voicing of the components to avoid unnatural onsets.

3 Packet Loss Concealment

The basic principle in the packet loss concealment method is to stretch the packets on each side of the missing packet interval, as illustrated in Figure 1. In this figure, S is the synthesis frame length when no packets are lost, which due to the 50% overlap is equal to half the analysis frame length. Δ_p and Δ_a are the additional lengths of the playout frames prior to and after the packet loss(es), respectively. We see that $\Delta_p + \Delta_a = K \cdot S$ where K is the number of consecutive packet losses. Note the difference in the analysis frame index m and synthesis frame index k as a consequence of lost packets not being given a synthesis index.

In the work presented here, we used $\Delta_a = \Delta_p$ but this could easily be relaxed. For example, if the packet after the loss interval is not yet present in the jitter buffer one could pick a large value for Δ_p and start playout of this packet and then calculate Δ_a when a packet arrives. Furthermore, if both packets are known it might be perceptually better to stretch one more than the other depending on the contents of the packets.

As indicated in Figure 1 the stretching of packets is carried out by modifying the point of time in which the amplitudes and frequencies of each packet occurs. This time-scale modification is carried out through a mix of parameter interpolation and

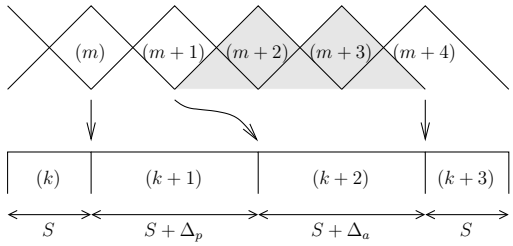

Figure 1: Principle for packet loss concealment scheme. Shaded frames symbolize packet losses.

overlap-add (OLA). Specifically, the l'th harmonic sinusoidal component is classified for interpolation or OLA by comparison to the corresponding harmonic from the previous synthesis frame. A component in the k'th frame is classified for interpolation if the following three conditions are met ($\hat{a}^{(k)}$ denotes the decoded model parameter a in the k'th frame):

- Both frequencies are below the voicing cut-off frequency of their respective frames, $l\hat{\omega}_0^{(k)} < \hat{\omega}_c^{(k)}$ and $l\hat{\omega}_0^{(k-1)} < \hat{\omega}_c^{(k-1)}$.
- The frequency difference is below 70 Hz, $|l\hat{f}_0^{(k)} - l\hat{f}_0^{(k-1)}| < 70$ Hz
- The amplitude ratio is below 5, $\max\left\{\frac{\hat{A}_l^{(k)}}{\hat{A}_l^{(k-1)}}, \frac{\hat{A}_l^{(k-1)}}{\hat{A}_l^{(k)}}\right\} < 5$

The first criterion means that unvoiced components will be overlap-added, where-as the other two prevent interpolation of dissimilar components. Note that unvoiced frames will be synthesized by OLA only.

3.1 Parameter Interpolation

For components matched by the three conditions above amplitudes are simply interpolated linearly over each synthesis frame, i.e. for $n = 0 \ldots S^{(k)} - 1$:

$$\tilde{A}_l^{(k)}(n) = \hat{A}_l^{(k-1)} + \frac{\hat{A}_l^{(k)} - \hat{A}_l^{(k-1)}}{S^{(k)}} n \tag{2}$$

Here $S^{(k)}$ denotes the length of the k'th synthesis frame. Likewise, frequencies evolve linearly over the frame, i.e. for $t \in [0, S^{(k)}]$:

$$\tilde{\omega}_l^{(k)}(t) = l\hat{\omega}_0^{(k-1)} + \frac{l\hat{\omega}_0^{(k)} - l\hat{\omega}_0^{(k-1)}}{S^{(k)}} t \tag{3}$$

From this continuous frequency model we determine the discrete phase function:

$$\begin{aligned}\tilde{\theta}_l^{(k)}(n) &= \tilde{\theta}_l^{(k)}(0) + \int_0^n \tilde{\omega}_l^{(k)}(t)dt \\ &= \tilde{\theta}_l^{(k)}(0) + l\hat{\omega}_0^{(k-1)}n + l\alpha^{(k)}n^2\end{aligned} \tag{4}$$

where $\alpha_l^{(k)} = \frac{1}{2}\left(\hat{\omega}_0^{(k)} - \hat{\omega}_0^{(k-1)}\right)/S^{(k)}$.

In order to avoid discontinuities at frame boundaries the initial phase $\tilde{\theta}_l^{(k)}(0)$ is the final phase of the same component in the previous frame:

$$\tilde{\theta}_l^{(k)}(0) = \tilde{\theta}_l^{(k-1)}\left(S^{(k-1)}\right) = \tilde{\theta}_l^{(k-1)}(0) + l\hat{\omega}_0^{(k-1)}S^{(k-1)} + l\alpha^{(k-1)}\left(S^{(k-1)}\right)^2 \tag{5}$$

That is, the initial phase is determined recursively from the pitch of previous synthesis frames back to the frame where the interpolation track was started.

3.2 Overlap-add Synthesis

The remaining, unmatched components are synthesized by OLA simply by stretching the overlap region of the analysis frames as sketched in Figure 2. However, after a packet loss the initial phases should be modified in order to compensate for the time offset Δ_a introduced here. Specifically:

$$\tilde{\phi}_l^{(k)} = \hat{\phi}_l^{(k)} - \Delta_a l\hat{\omega}_0^{(k)} \tag{6}$$

This step ensures that overlap-added components are properly matched to components synthesized by interpolation.

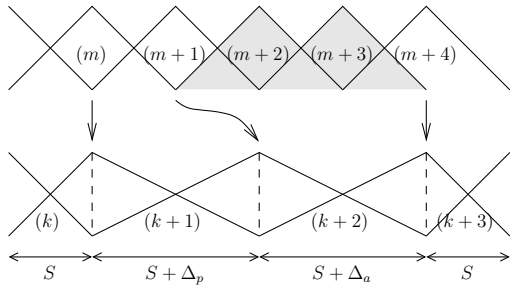

Figure 2: Overlap-add synthesis in case of packet loss. Shaded frames symbolize packet losses.

4 Experimental Results

In Figure 3 an example waveform resulting from the proposed method is shown for the case of 30 % random independent packet losses. We see that the structure of the missing parts is well synthesized. Simple listening tests have been carried out to investigate the

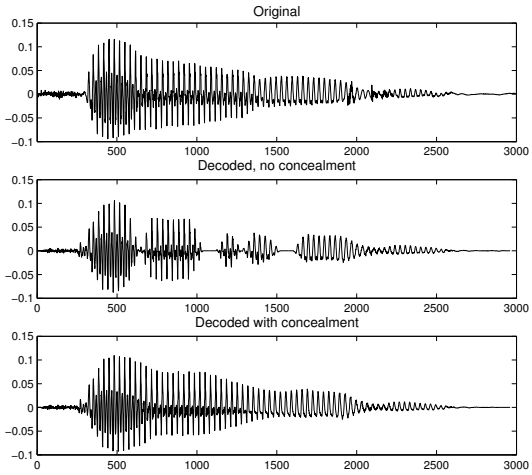

Figure 3: Example of packet loss concealment. The "Decoded, no concealment" sequence is obtained by silence substitution in lost frames.

performance of the method employed. The tests were conducted using a five point degradation score (Degradation Category Rating): degradation inaudible 5, audible but not annoying 4, slightly annoying 3, annoying 2, and very annoying 1 (see [13]). 12 untrained listeners participated. The test subjects were asked to grade the degradation of the signals relative to the original. Two test signals were used with each consisting of one speaker uttering one sentence. Three different realizations of four different cases of random independent packet losses were graded. In Table 2 the results of the listening tests are shown in the form of a mean score and a standard deviation for each test case. It can be seen that the average degradation due to the coding process has been

Packet loss	Mean Score	Std. Dev.
0%	3.8	0.9
10%	3.3	0.8
20%	3.2	0.9
30%	2.6	0.7

Table 2: Results of listening tests (mean score and standard deviation).

graded a little below 4 (audible but not annoying). The effectiveness of the proposed

packet loss concealment strategy is evident in that both the 10% and 20% packet loss cases are graded above 3 (slightly annoying), whereas the degradation in the 30% cases is more distinct and thus have received lower scores. These tests show that at average packet loss concealment can be successfully conducted in the compressed domain using the proposed methods. In fact, the degradation is only about 0.5 for packet losses of $10 - 20\%$ relative to the coded speech. The degradation is generally perceived as the synthesized speech becoming increasingly more tonal for higher packet losses. Also, the coded signal is slightly more tonal than the original.

5 Conclusion

In this paper a method for compressed domain packet loss concealment along with a sinusoidal speech coder for packet switched networks have been presented. The method has been evaluated by means of listening tests indicating that it reduces the consequences of packet losses with respect to perceived quality greatly. We therefore conclude that the combination of a sinusoidal coder and packet loss concealment operating on the compressed parameters provides an appealing solution for packet switched networks.

References

[1] D. J. Goodman, G. B. Lockhart, O. J. Wasem, and W. C. Wang, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(6), pp. 1440–1448, 1986.

[2] J. Lindblom, *Packetized Speech Transmission - Combatting the Packet Loss Problem*. Information Theory Group, School of EE&CE, Chalmers University of Technology, 2001.

[3] J. Lindblom and P. Hedelin, "Packet Loss Concealment Based on Sinusoidal Modeling," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, 2002, pp. 65–67.

[4] Y. J. Liang, N. Färber, and B. Girod, "Adaptive Playout Scheduling Using Time-Scale Modification in Packet Voice Communications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2001, pp. 1445–1448.

[5] F. Liu, J. Kim, and C.-C. J. Kuo, "Adaptive delay concealment for internet voice applications with packet based time-scale modification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2001, pp. 1461–1464.

[6] S. V. Andersen, W. B. Kleijn, R. Hagen, J. Linden, M. N. Murthi, and J. Skoglund, "ILBC - A Linear Predictive Coder with Robustness to Packet Losses," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, 2002, pp. 23–25.

[7] M. G. Christensen, C. Albøge, S. H. Jensen, and C. A. Rødbro, "A Harmonic Exponential Sinusoidal Speech Coder," in *Nordic Signal Processing Symposium*, 2002.

[8] C. A. Rødbro and S. H. Jensen, "Time-scaling of Sinusoids for Intelligent Jitter Buffer in Packet Based Telephony," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, 2002, pp. 71–73.

[9] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111(4), pp. 1917–1930, Apr. 2002.

[10] S. O. Rice, "Mathematical Analysis of Random Noise," *The Bell Systems Technical Journal*, vol. 3, pp. 282–332, 1944.

[11] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, 1991.

[12] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, pp. 3–14, 1993.

[13] *Revised Recommendation P.800 (Methods for Subjective Determination of Transmission Quality)*, ITU Std., Jan. 1996, COM 12-65-E.

Paper J

Subspace-based Fundamental Frequency Estimation

Mads Græsbøll Christensen, Søren Holdt Jensen, Søren Vang Andersen, and Andreas Jakobsson

The paper has been published in
The Proceedings of the XII. European Signal Processing Conference, pp. 637–640, 2004.

The layout has been revised.

Abstract

In this paper, we present a subspace-based fundamental frequency estimator based on an extension of the MUSIC spectral estimator. A noise subspace is obtained from the eigenvalue decomposition of the estimated sample covariance matrix and fundamental frequency candidates are selected as the frequencies where the harmonic signal subspace is closest to being orthogonal to the noise subspace. The performance of the proposed method is evaluated and compared to that of the non-linear least-squares (NLS) estimator and the corresponding Cramér-Rao bound; it is concluded that the proposed method has good statistical performance at a lower computational cost than the statistically efficient NLS estimator.

1 Introduction

The problem of estimating the fundamental frequency of a periodic signal is a classical problem in signal processing, and throughout the years many different solutions have been suggested to solve it. It is encountered in such applications as, for instance, coding of speech and audio, automatic music transcription and determination of rotating targets in radar. The problem of fundamental frequency estimation can be stated as follows; consider a harmonic signal with the fundamental frequency ω_0 that is corrupted by an additive white complex circularly symmetric Gaussian noise, $w(n)$, i.e.,

$$x(n) = \sum_{l=1}^{L} A_l e^{j(\omega_0 l n + \phi_l)} + w(n), \quad n = 0, \ldots, N-1 \tag{1}$$

where A_l and ϕ_l are the (real-valued) amplitude and the phase of the l'th harmonic, respectively. The problem considered in this paper amounts to estimating the fundamental frequency ω_0 from a set of N measured samples, $x(n)$. Note that the complex-valued signal model in (1) can also be applied to real-valued signals, when there is little or no spectral contents of interest in the frequencies near 0 and π, by the use of the discrete-time "analytical" signal [1]. The classical fundamental frequency estimators are typically time-domain techniques based on auto-correlation, cross-correlation, the average magnitude difference function (AMDF), or average squared difference function (ASDF). For a historical review of these methods, we refer to [2, 3], and for examples of more recent work we refer to [4–6]. While subspace techniques, such as the MUltiple SIgnal Classification (MUSIC) algorithm [7], have a rich history in spectral analysis in general, they have only rarely been used in fundamental frequency estimation. In [6, 8], MUSIC is used for finding the individual harmonics independently, and in [9], a noise estimate is obtained from MUSIC and used in a cepstral pitch estimator. In this paper, we propose an extension of the classical MUSIC algorithm by imposing the assumed harmonic structure in (1) on the MUSIC criterion. The paper is organized as follows. In Section 2, the covariance matrix model of the signal model (1) is presented along with

some definitions. Then, in Section 3, we present the proposed fundamental frequency estimator termed the harmonically constrained MUSIC estimator. In Section 4, some numerical results are presented and, finally, Section 5 concludes on the work.

2 Covariance Matrix Model

In this section, we present the covariance matrix model and introduce some useful vector and matrix definitions before we proceed to discuss the proposed extension. By assuming that the phases of the harmonics are independent and uniformly distributed in the interval $[-\pi, \pi]$, the covariance matrix $\mathbf{R} \in \mathbb{C}^{M \times M}$ can be written as [10]

$$\begin{aligned} \mathbf{R} &= \mathrm{E}\left\{\tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}^H(n)\right\} \\ &= \mathbf{A}(\omega_0)\mathbf{P}\mathbf{A}^H(\omega_0) + \sigma_w^2\mathbf{I}, \end{aligned} \tag{2}$$

where $\mathrm{E}\{\cdot\}$ denotes the statistical expectation, $(\cdot)^H$ the conjugate transpose, and $\tilde{\mathbf{x}}(n)$ is a signal vector containing M samples of the observed signal, i.e.,

$$\tilde{\mathbf{x}}(n) = \left[\begin{array}{cccc} x(n) & x(n-1) & \cdots & x(n-M+1) \end{array}\right]^T, \tag{3}$$

with $(\cdot)^T$ denoting the transpose. Further,

$$\mathbf{P} = \mathrm{diag}\left(\left[\begin{array}{ccc} A_1^2 & \cdots & A_L^2 \end{array}\right]\right) \tag{4}$$

and the full rank Vandermonde matrix $\mathbf{A}(\omega_0) \in \mathbb{C}^{M \times L}$ is defined as

$$\mathbf{A}(\omega_0) = \left[\begin{array}{ccc} \mathbf{a}(\omega_0) & \cdots & \mathbf{a}(\omega_0 L) \end{array}\right], \tag{5}$$

where

$$\mathbf{a}(\omega) = \left[\begin{array}{cccc} 1 & e^{-j\omega} & \cdots & e^{-j\omega(M-1)} \end{array}\right]^T. \tag{6}$$

Also, σ_w^2 denotes the variance of the additive noise, $w(n)$, and $\mathbf{I}$ is the $M \times M$ identity matrix. We note that

$$\mathrm{rank}\left(\mathbf{A}(\omega_0)\mathbf{P}\mathbf{A}^H(\omega_0)\right) = L, \tag{7}$$

and that the number of harmonics in $\mathbf{A}(\omega_0)$ is bounded by

$$L = \left\lfloor \frac{\omega_{max}}{\omega_0} \right\rfloor, \tag{8}$$

where ω_{max} may go up to π for real signals, although it is typically well below this. This is, for example, the case for audio sampled at 44.1 kHz or speech signals sampled at 16 kHz. Here, the constant $M \geq L + 1$ is a user parameter that determines the accuracy of the resulting MUSIC frequency estimator, with larger M yielding increasing resolution. Thus, M should be selected as large as possible while still allowing for a reliable estimate of the covariance matrix [10].

3 The Harmonic MUSIC Algorithm

The MUSIC algorithm [7, 11] (see also [12]) is based on the eigenvalue decomposition (EVD) of the covariance matrix $\mathbf{R}$, exploiting the structure in (2). Let

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H \tag{9}$$

where $\mathbf{U}$ is formed from the M orthonormal eigenvectors of $\mathbf{R}$, i.e.,

$$\mathbf{U} = \left[\begin{array}{ccc} \mathbf{u}_1 & \cdots & \mathbf{u}_M \end{array} \right], \tag{10}$$

and $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues, λ_k, on the diagonal. The following decomposition requires *a priori* knowledge of the number of harmonic components, L. Here, we will instead determine L using (8), and as a result this L will be frequency dependent; hereafter, we will use the notation $L(\omega_0)$ to stress this dependence. Now, let $\mathbf{G}(\omega_0)$ be formed from the $M - L(\omega_0)$ eigenvectors corresponding to the $M - L(\omega_0)$ least significant eigenvalues ($\mathbf{G}(\omega_0)$ is a function of ω_0 through $L(\omega_0)$), i.e.,

$$\mathbf{G}(\omega_0) = \left[\begin{array}{ccc} \mathbf{u}_{L(\omega_0)+1} & \cdots & \mathbf{u}_M \end{array} \right]. \tag{11}$$

Then, it can be shown that the noise subspace spanned by $\mathbf{G}(\omega_0)$ will be orthogonal to the Vandermonde matrix $\mathbf{A}(\omega_0)$ spanned by the L harmonic sinusoids [10], i.e.,

$$\mathbf{A}^H(\omega_0)\mathbf{G}(\omega_0) = \mathbf{0}. \tag{12}$$

We stress that where $\mathbf{A}$ is a function of the set of frequencies $\{\omega_l\}_{l=1}^{L}$ in MUSIC, it is here only a function of the fundamental frequency ω_0 as the frequencies of the harmonics are given by $\omega_l = \omega_0 l$. As $\mathbf{R}$ is typically unknown, one needs to form an estimate of it; here, we estimate the sample covariance matrix as

$$\widehat{\mathbf{R}} = \frac{1}{N}\sum_{n=M}^{N} \tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}^H(n). \tag{13}$$

and note that the orthogonality in (12) will only hold approximately for the eigenvectors found from this matrix. Exploiting the harmonic structure in (1), the estimated fundamental frequency can be found as

$$\arg\min_{\omega_0} \|\mathbf{A}^H(\omega_0)\mathbf{G}(\omega_0)\|_F, \tag{14}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. By the Cauchy-Schwarz inequality, we have that

$$\|\mathbf{A}^H(\omega_0)\mathbf{G}(\omega_0)\|_F \leq \|\mathbf{A}^H(\omega_0)\|_F \|\mathbf{G}(\omega_0)\|_F. \tag{15}$$

As the $M - L(\omega_0)$ columns of $\mathbf{G}(\omega_0)$ are orthonormal, and all the $L(\omega_0)$ columns of $\mathbf{A}(\omega_0)$ have norm $\sqrt{M}$, we get

$$\|\mathbf{A}^H(\omega_0)\mathbf{G}(\omega_0)\|_F \leq \sqrt{L(\omega_0)M}\sqrt{M - L(\omega_0)} \tag{16}$$

and thus

$$\frac{\|\mathbf{A}^H(\omega_0)\mathbf{G}(\omega_0)\|_F}{\sqrt{L(\omega_0)M(M-L(\omega_0))}} \leq 1. \tag{17}$$

We now define the harmonic pseudo-spectrum as

$$P(\omega_0) = \frac{L(\omega_0)M(M-L(\omega_0))}{\|\mathbf{A}^H(\omega_0)\mathbf{G}(\omega_0)\|_F^2}, \tag{18}$$

and find the estimated fundamental frequency as

$$\hat{\omega}_0 = \arg\max_{\omega_0 \in \Omega_0} P(\omega_0). \tag{19}$$

Thus, the fundamental frequency candidates can be found from (18) by sweeping ω_0 over a finite set of frequencies Ω_0 and then projecting the harmonic subspace onto the noise subspace. In the rest of this paper, we refer to this estimator as the *harmonically constrained* MUSIC or HMUSIC in short. The algorithm can be summarized as the following steps:

1. Estimate $\widehat{\mathbf{R}}$ using (13).
2. Perform an EVD of $\widehat{\mathbf{R}}$.
3. For each $\omega_0 \in \Omega_0$,
   (a) Determine $L(\omega_0)$ from (8).
   (b) Construct $\mathbf{A}(\omega_0)$ using (5), and $\mathbf{G}(\omega_\mathbf{0})$ using (11).
   (c) Compute $P(\omega_0)$ using (18).
4. Find fundamental frequency candidates as the maxima of $P(\omega_0)$.

We note that it is possible to use a noise subspace with a fixed dimension by estimating an upper bound on L. For example, in speech the fundamental frequency is typically limited to the range 60 Hz - 400 Hz, which would result in an upper bound of the dimension of the signal subspace

$$L = \left\lfloor \frac{\omega_{max} f_s}{2\pi 60} \right\rfloor, \tag{20}$$

with f_s being the sampling frequency. In our experience, the peaks of the harmonic pseudo-spectra computed using a fixed L are often more distinct and thus appear less noisy compared to the variable dimension approach, but the latter seems to give a better response at low frequencies.

A classical problem in fundamental frequency estimation is erroneous estimates at k or $1/k$ times the true fundamental frequency for $k = 2, 3, \ldots$, commonly referred to as doublings and halvings. These problems also exist in HMUSIC. Especially, doublings of the fundamental occur, because $\mathbf{A}(k\omega_0)$, for $k = 2, 3, \ldots$, is spanned by the column space of $\mathbf{A}(\omega_0)$ and these columns are thus also orthogonal to the noise subspace when this is kept fixed. This is less of a problem for variable dimension noise subspace. Halvings of the fundamental frequency also occur, but these are generally much weaker than the doublings as only a subset of the harmonics will be orthogonal to the noise subspace. In order to build a practical fundamental frequency estimator from HMUSIC, we need to limit the search space Ω_0 of ω_0 to some interval that does not include doublings and halvings. This can, for example, be achieved by pitch tracking, some coarse initial estimate, or by some post-processing of the harmonic pseudo-spectrum. In this paper, we defer from any further discussion of this and instead concentrate on the statistical performance of the estimator.

4 Experimental Results

4.1 Reference Methods

For reference, we use a non-linear least-squares (NLS) fundamental frequency estimator similar to that of [6]. This is a particularly simple version of the NLS frequency estimator (see, e.g., [10]) because of the harmonic relation between the sinusoidal components. As is well known, the NLS frequency estimator is statistically efficient under white noise conditions; furthermore, it can be shown that the NLS estimator is asymptotically efficient also for the coloured noise case [13]. We note that the NLS fundamental frequency estimator can be stated as the minimizer of the squared error between the signal and the harmonic sinusoidal model, and be found by sweeping over a finite set Ω_0 of frequencies. Here, we use the same grid as in HMUSIC. We refer to this method as harmonically constrained NLS (HNLS). While the HMUSIC gives strong false peaks at integer multiples of the fundamental frequency, the HNLS estimator is very prone to halvings ($1/k$ with $k = 2, 3, \ldots$) because a fundamental frequency of $0.5\omega_0$ will capture more signal energy than the true fundamental, especially under noisy conditions. Thus, like the HMUSIC, we need to limit the search range Ω_0 in order to get the correct result. For each grid point in Ω_0, HNLS is computationally more complex than HMUSIC as HNLS involves a matrix inversion and matrix products whereas HMUSIC involves only a matrix product for each frequency point. There is, however, some additional computational overhead associated with HMUSIC as it requires the calculation of the sample covariance matrix and an EVD. As the resolution of the grid increases, the relative influence of this overhead decreases. As an additional reference, we also use spectral MUSIC [7, 11] on the same grid as HMUSIC and HNLS to locate the frequency of the first harmonic. This method does not take the harmonic structure of the

spectrum into account.

4.2 Speech Signal

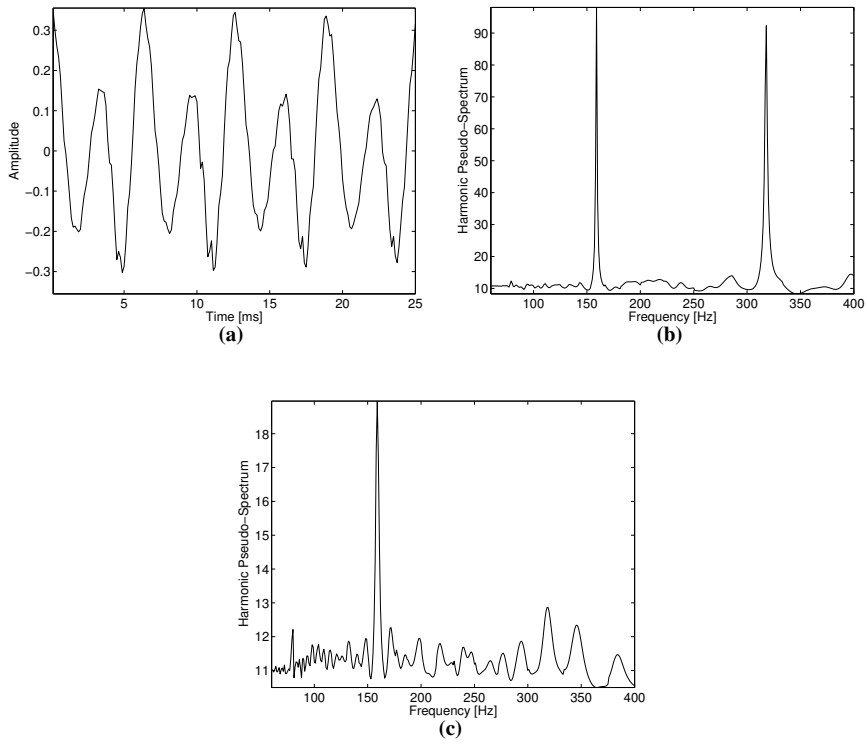

Figure 1: (a) Voiced speech segment, (b) harmonic pseudo-spectrum of speech segment using a fixed noise subspace, and (c) using a variable dimension subspace.

In this section, we show harmonic pseudo-spectra of a speech signal (female speaker, sampled at 8 kHz) and illustrate the difference between using a fixed dimensional noise subspace and a variable dimensional one. In Figure 1(b), the harmonic pseudo-spectrum of the segment of voiced speech in Figure 1(a) is depicted. This pseudo-spectrum has been calculated using a fixed noise subspace in the sweep over ω_0. It can be seen that the fundamental frequency stands out very clearly at approximately 159 Hz and that the double is very noticeably present at 318 Hz. As a comparison, the harmonic pseudo-spectrum with a variable dimension noise subspace is shown in Figure 1(c), clearly illustrating the reduced risk for a pitch-doubling. It can also be seen that the peaks of

Figure 1(b) are more distinct compared to the noise floor than those of 1(c).

4.3 Synthetic Signals

To investigate the statistical efficiency of HMUSIC, we perform an evaluation of the fundamental frequency estimator on a synthetic signal using a technique similar to those of [6, 14]. As a comparison, we also show the asymptotic Cramér-Rao bound (CRB) as derived in [14]. First, we investigate the effects of varying SNR for a fixed segment length. The SNR is defined as $SNR = 10\log_{10}(\sigma_s^2/\sigma_w^2)$, with σ_s^2 being the variance of the sinusoidal part of (1) and σ_w^2 being the variance of the noise. In Figure 2, the standard deviation of MUSIC, HNLS, HMUSIC and the CRB are shown as a function of the SNR for a segment length of 256 samples. These were found by 200 Monte Carlo simulations, where in each run the additive noise sequence and the phases of the harmonics have been randomized. A fundamental frequency of $\omega_0 = 2\pi 0.08$ (corresponding to 640 Hz at 8 kHz sampling frequency) was used in all simulations. Five real harmonics ($L = 10$) were used, all with an amplitude of 1, and the stepsize of the grid searches of MUSIC, HNLS and HMUSIC was set to 0.01 Hz. Further, Ω_0 was constrained to be in the vicinity of ω_0 by $\pm 10\%$, and M was set to 128.

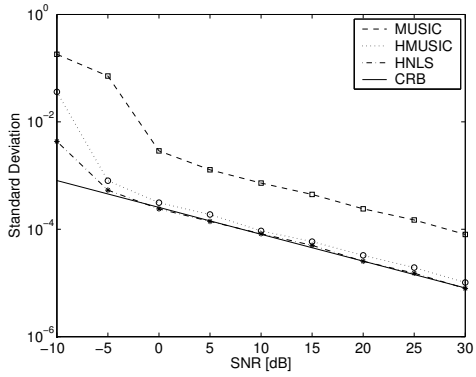

Figure 2: Standard deviation of the estimates $\hat{\omega}_0$ and the CRB for varying SNR for $N = 256$.

The effects of varying segment lengths, N, for a fixed SNR have also been investigated. The results are shown in Figure 3 for an SNR of 10 dB. Here a stepsize of 0.1 Hz was used and the dimensions of the sample covariance matrix was set to $M = \lfloor N/2 \rfloor$. Note that the HMUSIC and MUSIC algorithms are sensitive to the choice of M relative to N. From these figures, it can be seen that HMUSIC has very good statistical performance approaching the Cramér-Rao bound. From observing the performance of HMUSIC compared to MUSIC, it can also be seen that there is a big gain in taking the harmonic structure into account in the estimation.

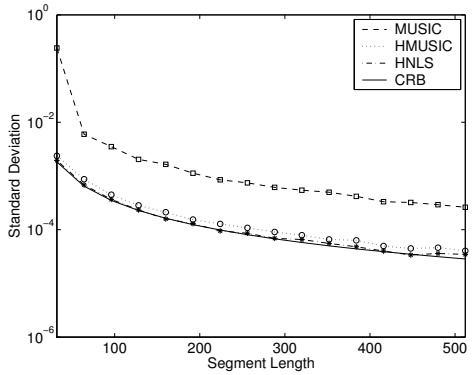

Figure 3: Standard deviation of the estimates $\hat{\omega}_0$ and the CRB for varying segment lengths N with $SNR = 10$dB.

5 Conclusion

In this paper, a subspace-based fundamental frequency estimator has been proposed. This estimator is based on a harmonic extension of the classical MUSIC estimator, letting the dimensionality of the noise signal subspace depend on the underlying fundamental frequency. The resulting estimator is obtained by sweeping over a set of frequencies. The performance of the estimator has been evaluated and compared to both the non-linear least-squares estimator, the classical MUSIC algorithm, and the Cramér-Rao bound. From the simulations, we conclude that the estimator has good statistical performance at a computational complexity, which is lower than the nonlinear least-squares for high resolutions.

References

[1] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Processing*, vol. 47, pp. 2600–2603, Sept. 1999.

[2] W. Hess, *Pitch Determination of Speech Signals*. Springer-Verlag, Berlin, 1983.

[3] W. Hess, "Pitch and voicing determination," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sohndi, Eds. Marcel Dekker, New York, 1992, pp. 3–48.

[4] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111(4), pp. 1917–1930, Apr. 2002.

[5] D. E. Terez, "Robust pitch determination using nonlinear state-space embedding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 345–348.

[6] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Processing*, vol. 80, pp. 1937–1944, 2000.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34(3), pp. 276–280, Mar. 1986.

[8] N. Malik and W. H. Holmes, "Pitch estimation and a measure of voicing from pseudo-spectra," in *Fifth International Symposium on Signal Processing and its Applications*, 1999.

[9] M. S. Andrews, J. Picone, and R. D. Degroat, "Robust pitch determination via SVD based cepstral methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1990, pp. 253–256.

[10] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Prentice Hall, 1997.

[11] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 306–309.

[12] G. H. Händel, "On the history of music," *IEEE Signal Processing Magazine*, p. 13, Mar. 1999.

[13] P. Stoica, A. Jakobsson, and J. Li, "Cisiod parameter estimation in the coloured noise case: Asymptotic cramér-rao bound, maximum likelihood, and nonlinear least-squares," in *IEEE Trans. Signal Processing*, vol. 45(8), Aug. 1997, pp. 2048–2059.

[14] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34(5), pp. 1124–1138, Oct. 1986.