JOINTLY OPTIMAL TIME SEGMENTATION, COMPONENT SELECTION AND QUANTIZATION FOR SINUSOIDAL CODING OF AUDIO AND SPEECH

Richard Heusdens and Jesper Jensen

Department of Mediamatics
Delft University of Technology
2628 CD Delft, The Netherlands

ABSTRACT

In this paper we propose a rate-distortion optimal algorithm for sinusoidal modeling of audio and speech. The algorithm determines for a pre-specified target bit-rate the optimal (variable-length) time segmentation, the optimal distribution of sinusoidal components over the segments and the optimal (scalar) quantizers for quantizing the sinusoid parameters. The optimization is done by jointly optimizing the segment lengths, number of sinusoids and quantizers using high-resolution quantization theory and dynamic programming techniques, which makes it possible to solve the algorithm in polynomial time. A particular advantage of the proposed method is that it solves the problem of, given a target bit-rate, finding the optimal balance between total number of sinusoids and number of bits per sinusoid.

1. INTRODUCTION

Sinusoidal modeling has proven to be an efficient technique for coding speech signals [1]. More recently, it has been shown that this method can also be exploited for low-rate audio coding [2, 3]. To account for the time-varying nature of the target signal, the sinusoidal analysis/synthesis is done on a (possibly variable-length) segment-by-segment basis, where each segment is modeled as a sum of sinusoids. After modeling, the sinusoid parameters are quantized and entropy coded.

The problem of optimally finding a variable-length segmentation and distributing a limited number of sinusoidal components over those segments has been studied in [4], where an algorithm has been presented that minimizes the total distortion subject to a bit rate constraint. This is done by jointly optimizing the segment lengths and number of sinusoids per segment using dynamic programming techniques, which makes it possible to solve the problem in polynomial time. With respect to quantization of the sinusoid parameters, optimal quantizers are presented in [5] (for amplitude and phase only) and [6] (amplitude, phase and frequency). These optimal quantizers, however, can only be computed if the sinusoids to be quantized are known.

Although the schemes mentioned above for quantization of sinusoid parameters and finding the segmentation and distribution of sinusoids over segments are individually optimal, they are not jointly optimal. Indeed, the scheme in [4] for finding an optimal segmentation and distribution of sinusoids requires knowledge of the cost in terms of rate and distortion of each and every candidate sinusoid. However, in order to determine this rate and distortion we need to know how the sinusoids are distributed. In other words, the optimal quantization and distribution of sinusoidal components depend on each other and need, therefore, to be optimized jointly.

In this paper we consider the joint optimization of time-segmentation, distribution of sinusoids and quantization of sinusoid parameters. We restrict ourselves to amplitude and phase quantization only, although the incorporation of frequency quantization is straightforward once the optimal frequency quantizers are known. To account for human auditory perception and variable-length segments, we define an appropriate norm on the signal space that depends on both the analysis window and the masking threshold [7]. Similar to the algorithm proposed in [4], the algorithm presented here has a searching complexity of only $\mathcal{O}(M^2)$, where M is the total number of allowed segment boundaries throughout the signal, rather than $\mathcal{O}(2^M)$ which would be needed for an exhaustive search.

2. HIGH-RESOLUTION QUANTIZATION THEORY

In order to minimize the quantization distortion subject to an entropy constraint, we have to define a proper distortion measure. To do so, we define a norm on our signal space which incorporates both perception and the effect of windowing as

$$\|x\|^2 = \int_0^1 \hat{a}(f)|(\hat{wx})(f)|^2 df, \qquad (1)$$

where $\hat{}$ indicates the Fourier transform operation, w is a window defining the signal segment, and $\hat{a}$ is a weighting function representing the sensitivity of the human auditory system which we select to be the inverse of the masking threshold [7]. By doing so, regions in which the auditory system is less sensitive will contribute less to the total distortion as compared to regions in which the auditory system is more sensitive.

Quantization distortion

In order to determine the quantization distortion, we will first consider the case where the input signal consists of

The research was conducted within the ARDOR project, supported by E.U. grant no. IST-2001-34095.

a single complex exponential, and later generalize this to the practically more relevant case where the input signal consists of several sinusoids. To do so, let $x = ae^{i2\pi\phi}e^{i2\pi\nu(\cdot)}$ be the input signal, $a \geq 0$ and $\phi, \nu \in [0, 1)$, and let $\tilde{x} = \tilde{a}e^{i2\pi\tilde{\phi}}e^{i2\pi\nu(\cdot)}$ denote the quantized version of x. Given $\tilde{x}$, we can compute the quantization distortion, which is given by

$$\begin{aligned} d(x, \tilde{x}) &= \|x - \tilde{x}\|^2 \\ &= \int \hat{a}(f)|ae^{i2\pi\phi} - \tilde{a}e^{i2\pi\tilde{\phi}}|^2|\hat{w}(f-\nu)|^2 df \\ &= c_\nu |ae^{i2\pi\phi} - \tilde{a}e^{i2\pi\tilde{\phi}}|^2, \end{aligned}$$

where $c_\nu = \int \hat{a}(f)|\hat{w}(f-\nu)|^2 df = \|e^{i2\pi\nu(\cdot)}\|^2 > 0$. Hence, the (perceptual) distortion is given by the squared difference between the original and quantized complex amplitude, multiplied by a constant c_ν which depends on both the analysis window and the masking threshold.

In this paper, we define our amplitude and phase quantizer as consisting of a doubly indexed set of cells $\mathcal{C} = \{c_{k,l} : k, l \in \mathbb{Z}\}$ together with a corresponding set of reproduction points $\mathcal{R} = \{r_{k,l} = \tilde{a}_{k,l}e^{i2\pi\tilde{\phi}_{k,l}} : k, l \in \mathbb{Z}\}$. The expected distortion, say D, can then be computed as the average distortion over all quantization cells, that is,

$$\begin{aligned} D(x, \tilde{x}) &= E\|x - \tilde{x}\|^2 \\ &= c_\nu \sum_k \sum_l \iint_{c_{k,l}} f_{A,\Phi}(a, \phi)|ae^{i2\pi\phi} - \tilde{a}_{k,l}e^{i2\pi\tilde{\phi}_{k,l}}|^2 da d\phi. \end{aligned}$$

If we assume that the amplitude and phase quantizer step sizes, $\Delta_{a_{k,l}}$ and $\Delta_{\phi_{k,l}}$ respectively, are sufficiently small, the pdf $f_{A,\Phi}$ is approximately constant within each quantization cell (high-resolution assumption), the expected distortion D may be approximated by

$$\begin{aligned} D(x, \tilde{x}) \approx \\ c_\nu \iint f_{A,\Phi}(a, \phi) \left(\frac{g_A^{-2}(a, \phi)}{12} + a^2 \frac{g_\Phi^{-2}(a, \phi)}{12} \right) da d\phi, \end{aligned}$$

where g_A and g_Φ are the (unnormalized) quantizer point densities [8, 9] which when integrated over a region S gives the total number of quantization levels within S. In the case of one-dimensional quantizers, this means that the quantizer step sizes are simply given by the reciprocal values of g_A and g_Φ evaluated at a and ϕ, respectively.

In the case of sinusoidal modeling, the input signal generally consists of multiple, say L, exponentials, that is,

$$x = \sum_{l=1}^{L} a_l e^{i2\pi\phi_l} e^{i2\pi\nu_l(\cdot)}.$$

In this case, the total quantization distortion consists of the quantization distortion of the individual components plus a contribution due to the mutual interaction of the components. The mutual interaction, however, can in most practical situations be neglected [10] so that the total distortion simply becomes

$$\begin{aligned} D(x, \tilde{x}) \approx \sum_{l=1}^{L} c_{\nu_l} \iint f_{A_l,\Phi_l}(a, \phi) \cdot \\ \left(\frac{g_{A_l}^{-2}(a, \phi)}{12} + a^2 \frac{g_{\Phi_l}^{-2}(a, \phi)}{12} \right) da d\phi, \quad (2) \end{aligned}$$

where we have introduced the subscript l for reference to the lth sinusoid.

Entropy

In order to compute the entropy H of the reproduction points $\mathcal{R}$, let $p_{k,l} = P(r_{k,l}) = P(ae^{i2\pi\phi} \in c_{k,l})$ denote the probability that the complex amplitude $ae^{i2\pi\phi}$ lies in cell $c_{k,l}$. Under high-resolution assumptions, the probabilities $p_{k,l}$ can be approximated by $p_{k,l} \approx f_{A,\Phi}(\tilde{a}_{k,l}, \tilde{\phi}_{k,l})\Delta_{a_{k,l}}\Delta_{\phi_{k,l}}$ so that, assuming that entropies are additive over sinusoidal components[1], the total entropy of the sinusoid parameters becomes

$$\begin{aligned} H = \sum_{l=1}^{L} \Bigg(h(A_l, \Phi_l) + \iint f_{A_l,\Phi_l}(a, \phi) \log(g_{A_l}(a, \phi)) da d\phi \\ + \iint f_{A_l,\Phi_l}(a, \phi) \log(g_{\Phi_l}(a, \phi)) da d\phi \Bigg), \quad (3) \end{aligned}$$

where $h(A_l, \Phi_l)$ is the differential entropy of the source variables (A_l, Φ_l).

3. PROBLEM STATEMENT

As mentioned before, with sinusoidal modeling the analysis/synthesis of the input signal is done on a segment-by-segment basis, with each segment being modeled as a sum of complex exponentials. The number of sinusoidal components per segment is finite so that, in general, the modeled signal is only an approximation of the input signal. Clearly, this modeling error, which we will denote by $D^{(m)}$, depends on both the segmentation and the number of components used. After quantizing the sinusoid parameters, the total distortion introduced is thus a combination of modeling and quantization distortion, the latter denoted by $D^{(q)}$. Under high-resolution assumptions, the quantization error behaves much like random noise (it has small correlation with the modeled signal and has approximately a flat spectrum), leading to an additive-noise model of quantization noise. Consequently, we assume that the total distortion introduced by modeling and quantization can be approximated by the sum of the two distortions.

The problem we consider in this paper is how to find the time-segmentation and given this segmentation, the distribution of (quantized) sinusoidal components that minimizes the total distortion (modeling and quantization distortion) such that the associated number of bits needed to uniquely decode the quantized signal $\tilde{x}$ does not exceed a predefined target rate, say R_t. In order to solve

[1] In the work presented here we restrict ourselves to independent coding of the sinusoid parameters and do not exploit the fact that in practical situations redundancy between parameters can be removed.

this problem we need to formalize it. To do so, let $s = \{s_1, s_2, \ldots, s_K\}$ denote a particular time segmentation of the input signal consisting of, possibly variable-length, disjoint segments s_k, where each s_k is restricted to be an integer multiple of a predefined length. Moreover, denote by $c_s = \{L_1, L_2, \ldots, L_K\}$ a particular distribution of sinusoidal components over the segmentation s, where L_k denotes the number of components assigned to segment s_k. In addition, let $g_{A,s} = \{g_{A,s_1}, g_{A,s_2}, \ldots, g_{A,s_K}\}$ and $g_{\Phi,s} = \{g_{\Phi,s_1}, g_{\Phi,s_2}, \ldots, g_{\Phi,s_K}\}$, where $g_{A,s_k} = \{g_{A_{k,1}}, g_{A_{k,2}}, \ldots, g_{A_{k,L_k}}\}$ and $g_{\Phi,s_k} = \{g_{\Phi_{k,1}}, g_{\Phi_{k,2}}, \ldots, g_{\Phi_{k,L_k}}\}$ denote a particular set of quantizer point densities for amplitudes and phases, respectively, for segment s_k. With these definitions, the formal problem statement can be written as

$$\min_{s} \min_{c_s} \min_{g_{A,s}} \min_{g_{\Phi,s}} \left(D^{(m)} + D^{(q)} \right) \quad \text{subject to } H \leq R_t. \tag{4}$$

The standard approach to solve the constrained problem (4) is to introduce a Lagrange multiplier $\lambda > 0$ and to minimize the Lagrangian cost functional $J(\lambda) = D^{(m)} + D^{(q)} + \lambda H$ where λ should be chosen such that the entropy constraint $H = R_t$ is met. The problem formulation in (4), however, is an NP hard problem since the searching complexity for finding the optimal segmentation is $\mathcal{O}(2^M)$, where M is the total number of allowed segment boundaries throughout the signal. In order to solve (4) in polynomial time we, therefore, follow the approach in [4] and assume that entropies and distortions are additive and independent over segments, that is,

$$D^{(m)} = \sum_{k=1}^{K} D_k^{(m)}, \quad D^{(q)} = \sum_{k=1}^{K} D_k^{(q)} \quad \text{and} \quad H = \sum_{k=1}^{K} H_k,$$

with $D_k^{(q)}$ and H_k given by (2) and (3), respectively. With these definitions, the total Lagrangian cost functional $J(\lambda)$ is additive over segments as well. Hence, since entropies and distortions are assumed to be independent over segments and $D^{(m)}$ depends only on s and c_s but *not* on $g_{A,s}$ and $g_{\Phi,s}$, we can write our minimization problem as

$$\min_{s} \min_{c_s} \min_{g_{A,s}} \min_{g_{\Phi,s}} J(\lambda) = \min_{s} \sum_{k=1}^{K} \min_{L_k} \left(D_k^{(m)} + \min_{g_{A,s_k}} \min_{g_{\Phi,s_k}} \left(D_k^{(q)} + \lambda H_k \right) \right). \tag{5}$$

In words, (5) says that we first have to find, for each and every segment s_k, the optimal quantizer point densities $g_{A_{k,l}}$ and $g_{\Phi_{k,l}}$, next the optimal number of components L_k, and finally the best segmentation s. After this, however, we have to determine the optimal value of λ which, as we will show in Section 5, is a convex optimization problem which can be solved using standard techniques like the bisection method or Newton's method. Note that the right-hand side of (5) describes a minimization of an additive sum of independent terms, which suggests to use the approach of dynamic programming. By doing so, the searching complexity for the best segmentation is only $\mathcal{O}(M^2)$, rather than $\mathcal{O}(2^M)$ which would be needed for an exhaustive search.

4. OPTIMAL QUANTIZATION POINT DENSITIES

By inspection of (5), we conclude that the optimal quantization point densities are found by solving

$$\min_{g_{A,s_k}} \min_{g_{\Phi,s_k}} \left(D_k^{(q)} + \lambda H_k \right), \tag{6}$$

for all s_k. The optimal densities are found using elementary calculus of variations [11], yielding

$$g_{A_{k,l}} = \left(\frac{c_{\nu_{k,l}}}{6\lambda \log(e)} \right)^{\frac{1}{2}}, \tag{7}$$

and

$$g_{\Phi_{k,l}} = \left(\frac{a^2 c_{\nu_{k,l}}}{6\lambda \log(e)} \right)^{\frac{1}{2}} = a g_{A_{k,l}}, \tag{8}$$

where the indices $k = 1, \ldots, K$ and $l = 1, \ldots, L_k$ refer to a particular segment and component within that segment, respectively. In the case that the distribution of the sinusoids is known, the optimal λ is found by substitution of (7) and (8) into (3) for each and every segment s_k and equate the total sum over all segments to R_t, the total target rate. However, in the optimization problem at hand, this distribution is not known in advance so that we cannot analytically determine the optimal λ and thus the optimal point densities. We can, however, overcome this problem, as we will show in the next section, by iteratively finding the optimal λ.

5. FINDING THE OPTIMAL λ

Let us assume that we are given a particular value of λ. Given λ, we can compute the quantization point densities $g_{A_{k,l}}$ and $g_{\Phi_{k,l}}$ which are given by (7) and (8), respectively. These densities, however, are optimal with respect to a target rate $R_t' \neq R_t$, unless λ is optimal. Clearly, in that case the entropy and quantization distortion can be calculated for each and every possible sinusoidal component and is given by

$$H_{k,l} = h(A_{k,l}, \Phi_{k,l}) + b(A_{k,l}) + \log(c_{\nu_{k,l}}) - \log(6\lambda \log(e)),$$

and

$$D_{k,l} = \lambda \log(e),$$

respectively, where $b(A_{k,l}) = \int f_{A_{k,l}}(a) \log(a) da$. Note that the quantization distortion in this case is independent of k and l. Hence, in the optimal case, all sinusoidal components contribute equally to the total distortion, as is to be expected since the norm defined in (1) is a weighted squared-error distortion measure. Hence, for a given value of λ, we are able to solve (5), resulting in a particular segmentation, distribution of sinusoids and corresponding optimal quantizers. The associated bit rate is given by

$$R_t' = H(\lambda) = \sum_{k=1}^{p} \sum_{l=1}^{L_k} \left(h(A_{k,l}, \Phi_{k,l}) + b(A_{k,l}) + \log(c_{\nu_{k,l}}) - \log(6\lambda \log(e)) \right).$$

In the case that $H(\lambda) \neq R_t$, we can modify λ and determine the optimal densities for the modified λ and repeat this procedure until the target rate R_t is met.

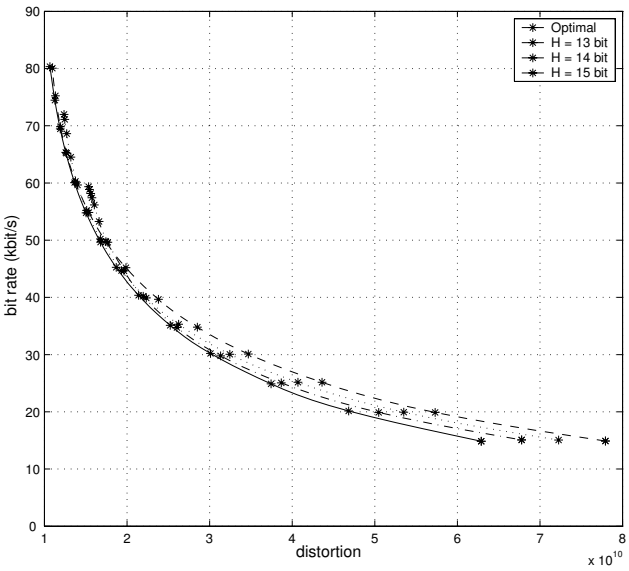

Figure 1: Rate-distortion curves for both the optimal situation and the situation of a fixed number of bits per components.

6. EXPERIMENTAL RESULTS

In this section we discuss experimental results obtained by computer simulations. The test excerpt used for the experiments is 48 kHz sampled contemporary pop music. Sinusoidal components were extracted using the psychoacoustical matching pursuit algorithm [12] with the perceptual distortion measure as presented in [7]. We computed the optimal time segmentation, distribution of sinusoids and quantizers using the theory developed above and compared these results to the situation where quantizers were determined independent of the time segmentation and distribution of sinusoidal components. In the latter case, segmentation and distribution of sinusoids is computed using the algorithm described in [4] for a fixed entropy, say H bit, per sinusoid. Subsequently, optimal amplitude and phase quantizers are computed using high-resolution quantization theory [5, 6]. We shall refer to this scheme as the two-stage approach. Figure 1 shows rate-distortion curves for different values of the entropy H. As can be seen from the figure, the optimal choice for H in the two-stage approach (right most curves) depends on the target bit rate; for low target rates (e.g. 15 kbit/s) the choice $H = 13$ outperforms the choice $H = 15$, whereas at higher rates (e.g. 50 kbit/s) the opposite is true. Which value to choose for H, however, is not straightforward and must in general be determined by trying out different values. This (exhaustive search) can be overcome by jointly optimizing segmentation, distribution and quantization, as shown in Figure 1 (solid line). It should be noted that even in the case where we are willing to try out many H-values, the performance will always be less than that of the algorithm proposed here since in the latter case the two stages are optimized jointly.

7. REFERENCES

[1] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on ASSP*, vol. 34, no. 4, pp. 744–754, August 1986.

[2] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC - analysis/synthesis codec for very low bit rates," in *Preprint 4179 (F-6) 100th AES Convention*, 1996.

[3] K.N. Hamdy, M. Ali, and A.H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representation," in *Proceedings ICASSP'96*, 1996, pp. 1045–1048.

[4] Z. Xiong, K. Ramchandran, C. Herley, and M.T. Orchard, "Flexible tree-structured signal expansions using time-varying wavelet packets," *IEEE Trans. on Signal Processing*, vol. 45, no. 2, pp. 333–345, February 1997.

[5] R.Vafin and W.B. Kleijn, "Entropy constrained quantization using unrestricted polar quantizers," *IEEE Trans. on Speech and Audio Processing*, 2004, Accepted for publication.

[6] P.E.L. Korten, R. Heusdens, and J. Jensen, "High rate spherical quantisation of sinusoidal parameters," in *Proceedings 12th European Signal Processing Conference*, Vienna, Austria, September 6-10 2004, pp. 1757–1760.

[7] R. Heusdens and S. van der Par, "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits," in *Proceedings ICASSP 2002*, Orlando, Florida, USA, May 13-17 2002, pp. 1809–1812.

[8] S.P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. on Information Theory*, vol. 28, pp. 129–137, 1982.

[9] R.M. Gray and D.L. Neuhoff, "Quantization," *IEEE Trans. on Information Theory*, vol. 44, no. 6, pp. 2325–2383, October 1998.

[10] R.P. Westerlaken, "High-resolution quantisation of sinusoidal parameters using a perceptual distortion measure," M.S. thesis, Delft University of Technology, Delft, The Netherlands, June 2004, Techical Report TR-200405.

[11] Hans Sagan, *Introduction to the Calculus of Variations*, Dover Books on Mathematics. Dover Publications, Inc., New York, 1969.

[12] R. Heusdens, R. Vafin, and W.B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Letters*, vol. 9, no. 8, pp. 262–265, August 2002.