TIME-DIFFERENTIAL ENCODING OF SINUSOIDAL MODEL PARAMETERS FOR MULTIPLE SUCCESSIVE SEGMENTS

Jesper Jensen and Richard Heusdens

Dept. of Mediamatics
Delft University of Technology
Delft, The Netherlands
E-mail: {J.Jensen, R.Heusdens}@ewi.tudelft.nl

ABSTRACT

Sinusoidal coding is a key technique for low rate audio coding. In sinusoidal coding, the target signal is represented by perceptually relevant sinusoids; however, often the sinusoids are estimated without taking into account that the sinusoidal parameters are going to be differentially encoded. In this paper we present an algorithm for joint extraction and time-differential encoding of sinusoidal model parameters. For a pre-specified target bit rate, the algorithm extracts the set of sinusoids for a sequence of signal segments which lead to minimum distortion in the reconstructed signal. Furthermore, it determines which sinusoids in a given segment should be encoded time-differentially relative to which in the previous segment. Simulation experiments show that the proposed algorithm leads to a reduction of 3-5% in bit rate compared to a state-of-the-art time-differential sinusoidal coding system.

1. INTRODUCTION

In low bit-rate audio compression systems, the target signal to be encoded is typically represented using a set of complementary signal models. Often, the model set includes sinusoidal, noise, and transient models [1, 2]. On the encoder side, model parameters are estimated, quantized, encoded and transmitted to the decoder, where the decoded parameter set is used for reconstructing the quantized target signal.

Practically all low bit-rate audio coders employ a sinusoidal model for representing the periodic constituents of the target signal [1, 2, 3, 4]. Traditionally, the sinusoidal model represents signal segments as linear combinations of sinusoidal functions, each represented by an amplitude, a frequency, and possibly a phase parameter. In order to minimize the bit-rate needed for representing the sinusoidal parameters, inter- and/or intra-segment parameter correlations may be exploited using time-differential (TD) or frequency-differential (FD) encoding techniques, see e.g. [3] and [5], respectively. We focus in this paper on TD-techniques.

Fig. 1 shows traditional blocks in a TD-based sinusoidal encoding system, see e.g. [1]. The target signal is divided into consecutive segments of suitable durations, and perceptually relevant sinusoids are estimated for each segment. In TD-based encoding schemes, the inter-segment correlations between sinusoidal components are exploited by associating components in a given segment with components in the previous segment (so-called linking), thereby forming 'tracks' in the time/frequency plane. Then, short or perceptually less relevant tracks may be eliminated [1], and the parameters of the remaining tracks quantized, according to a given available target bit budget. Finally, the quantizer indeces are encoded and packed to form a bitstream. Although the problems solved by the blocks in Fig. 1 depend on each other, they are typically solved one at a time, without taking into account this dependency (although some coders employ iterations over some of the blocks). For example, often, the 'Parameter Estimation' block extracts the set of perceptually most relevant sinusoids for a given segment, without taking into account their associated bit rate.

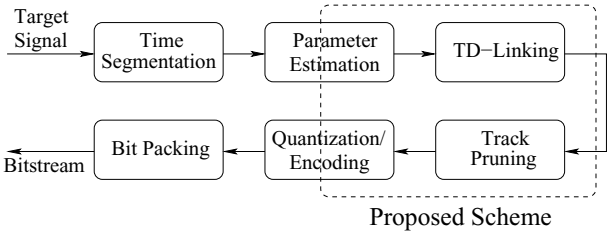

Figure 1: Block diagram of traditional TD-based sinusoidal coder (the block order of a specific coder may differ from the one shown here).

In this paper we describe an algorithm which does take into account the inter-dependence between several of the blocks in Fig. 1: it aims at finding a jointly optimal solution of the subproblems indicated by the dashed box in Fig. 1. The algorithm operates in a rate-distortion (R-D) framework, where the distortion in the reconstructed target signal is minimized subject to a bit rate constraint. Specifically, the algorithm finds for a sequence of signal segments the sinusoids which when TD-quantized and encoded at a certain pre-specified bit rate results in minimum distortion in the reconstructed signal. Additionally, it determines which sinusoids in a given segment should be quantized/encoded using TD-techniques and which should be quantized directly, i.e., without using differential techniques. Finally, for TD-encoded sinusoids the algorithm determines which sinusoidal components in a given segment are matched to which in the previous segment.

The algorithm proposed here extends on the one presented in [6] in one important way. In [6] sinusoidal components are distributed across segments without taking into account the fact that some of them may be TD-encoded, and, for this reason, be 'cheaper' in terms of bit rate than others. The proposed algorithm does take this fact into account. Specifically, it distributes sinusoids jointly over several consecutive segments and determines the optimal sequence of TD-relations between the segments.

2. OPTIMAL TD-ENCODING OF MULTIPLE SEGMENTS

Let us consider the case where the target signal has been divided into consecutive segments, and assume that for each such segment a number of *candidate* sinusoidal components have been estimated. Furthermore, assume that the quantizers needed for quantizing the corresponding model parameters are given. A segment must be represented by a subset of its candidate components, and we allow each of these to be quantized and encoded either directly, i.e., without differential techniques, or differentially relative to one of the components in the previous segment. For the differential case we impose a similar constraint as in [6] that no two components in a segment can be quantized and encoded relative to the same component

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

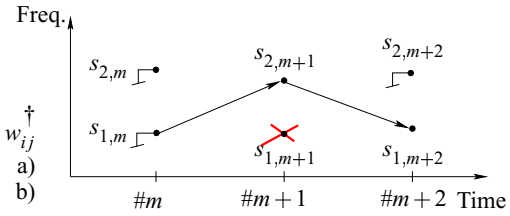

Figure 2: Example of coding template sequence. Black dots represent candidate components. Arrows and ground symbols indicate TD and direct quantization, respectively. Crosses indicate discarded candidate components.

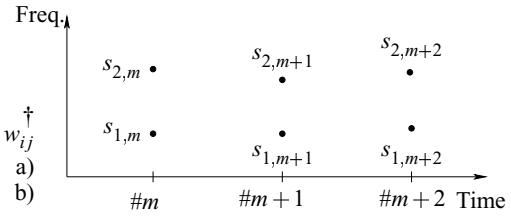

Figure 3: Small-scale example of super-segment containing three segments with two candidate components in each.

in the previous segment.

For a sequence of segments we wish to answer the following questions: a) which candidate components should be used (and which should be discarded), b) which of the used components should be TD-encoded and which should be encoded directly, and c) which of the TD-encoded components should be matched to which in the previous segment. The problem at hand is to answer these questions such that the resulting distortion in the reconstructed signal is minimized under a bit rate (or entropy) constraint.

Let us consider the consequences of some of the answers to questions a)–c) in terms of rate and distortion. If a candidate component is discarded, it will not appear in the sinusoidal reconstruction, and, consequently, a *modeling error* will occur. If it is chosen to use a candidate in the reconstruction it may be quantized directly or differentially. In either case a (typically smaller) *quantization error* occurs; however, in both these cases the smaller error is achieved at the expense of a certain bit cost.

We consider the situation where consecutive segments of the target signal have been grouped into L non-overlapping 'super-segments', and where selected candidate components in the first segment of each super-segment are restricted to be quantized directly. This situation is of interest for packet based transmission channels, because it allows a given super-segment to be decoded even if the packet containing the previous super-segment did not arrive at the decoder.

In order to formulate our problem, we introduce the concept of *coding templates*. For a given segment, a coding template represents one specific combination of answers to the questions a)–c) posed above. Consider for example segment $m+1$ in the small-scale example of Fig. 2; in this example the coding template represents the case where the first candidate component is discarded, while the second one is quantized differentially relative to the first component in the previous segment.

2.1 Problem Formulation

Our goal is find coding template *sequences* for each super-segment such that the total per signal distortion is minimized subject to a bit-rate constraint. Assuming that distortions and rates are additive across super-segments, our problem can be formulated as follows:

$$\min_{v=[v_1\ v_2\ \cdots\ v_L]} \sum_{l=1}^{L} D_l(v_l) \text{ such that } \sum_{l=1}^{L} R_l(v_l) \leq R_t, \qquad (1)$$

where v_l denotes a coding template sequence for super-segment l, D_l and R_l denotes the distortion and rate, respectively, in super-segment l, and R_t denotes the total target bit budget. Fig. 2 shows a small-scale example of a particular coding template sequence for a super-segment consisting of three segments with two candidates in each segment. Using the method of Lagrange multipliers and observing that super-segments can be treated independently, the problem in Eq. (1) can be solved by minimizing the following lagrangian cost function:

$$\sum_{l=1}^{L} \min_{v_l} \left(D_l(v_l) + \lambda R_l(v_l) \right), \qquad (2)$$

where $\lambda \geq 0$ is a Lagrange multiplier. We note that Eq. (2) consists of L independent minimization problems, one for each super-segment. Assuming that distortions and rates are additive across segments, we can formulate the subproblems of Eq. (2) as follows

$$\min_{v_l=[v_{1,l}\ v_{2,l}\ \cdots\ v_{M_l,l}]} \sum_{m=1}^{M_l} \left(D_{m,l}(v_{m,l}) + \lambda R_{m,l}(v_{m,l}) \right), \qquad (3)$$

where M_l is the number of segments in super-segment l, $v_{m,l}$ denotes the coding template of the m'th segment in super-segment l, and $D_{m,l}$ and $R_{m,l}$ denote the corresponding distortion and rate, respectively. We note that the coding templates in a sequence $v_l = [v_{1,l}\ v_{2,l}\ \cdots\ v_{M_l,l}]$ are interdependent. Choosing e.g. template $v_{m-1,l}$ in segment $m-1$ will affect the set of valid coding templates for segment m. Taking these coding template interdependencies into account when solving the subproblem in Eq. (3) complicates the solution.

2.2 Problem Solution

In the following we focus on a solution of the subproblem in Eq. (3), which takes into account the interdependencies mentioned above. Since this problem occurs locally within a super-segment, we shall, for notational convenience, omit the super-segment subscript.

We represent the set of valid coding template sequences in Eq. (3) using a bipartite graph. Consider for illustration purposes the small-scale example shown in Fig. 3; the super-segment shown here consists of three segments each containing two candidate components. It can be verified that the set of valid coding template sequences for this example can be represented by the bipartite graph shown in Fig. 4. The graph consists of three subgraphs (of which one has been marked by a dashed box) which represent the set of coding templates for each segment. The top subgraph is different from the others, because candidate components in the first segment of a super-segment are restricted to be directly encoded; this point will become clear from what follows.

Consider here the subgraph for segment $m+1$ indicated by the dashed box in Fig. 4. The subgraph consists of two disjoint set of nodes, one set (on the left-hand side) related to the previous segment (m) and one set (one the right-hand side) related to segment $m+1$. Denote by K_m the number of candidate components in segment m. The right-hand node set contains K_{m+1} nodes $s_{1,m+1} \cdots s_{K_{m+1},m+1}$ which represent the candidate components, and additionally $2K_{m+1}$ dummy nodes. Edges between the candidate nodes on the right-hand side and the left-hand node set represent possible consequenses for the candidate components. Specifically, edges between $s_{i,m}$ and $s_{j,m+1}$ nodes represent differential quantization and encoding of candidate component number j relative to component i in the previous segment. Edges between 'ground' and $s_{j,m+1}$ nodes represent direct quantization of candidate component $s_{j,m+1}$, while edges between † and $s_{j,m+1}$ nodes correspond to the case where candidate component $s_{j,m+1}$ is discarded. The edges between the top-right nodes in one subgraph and the bottom-left nodes of the following are important in the sense that they ensure that can-

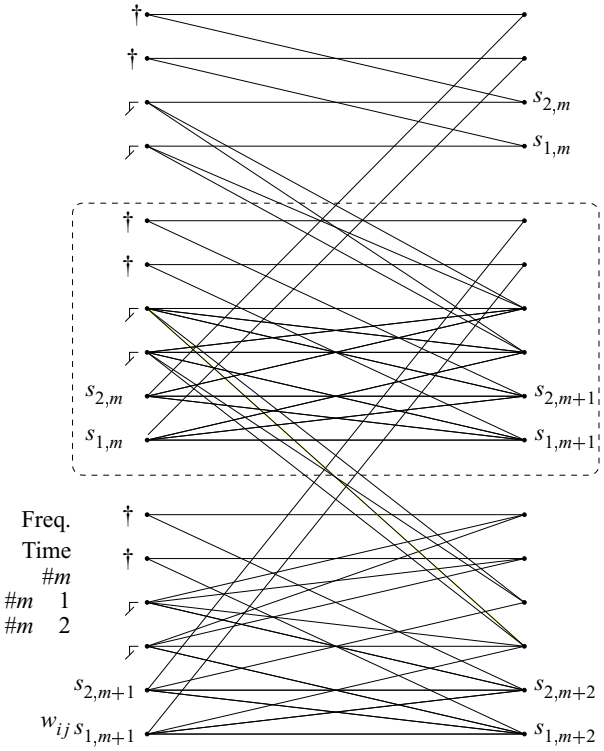

Figure 4: Bipartite graph for representing the set of valid coding template sequences for the case of three consecutive segments each containing two candidate components.

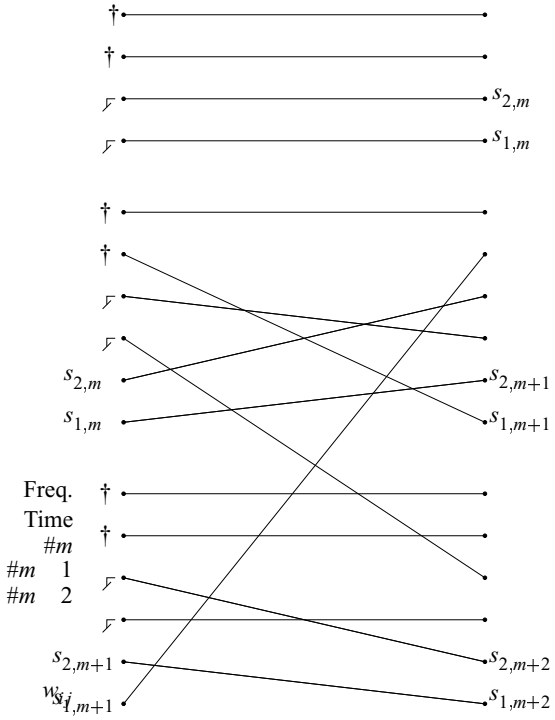

Figure 5: Example of linear assignment in bipartite graph corresponding to coding template sequence shown in Fig. 2.

didates discarded in one segment are not used in TD-relations in the following segment.

It can be verified that each valid coding template sequence corresponds to a *linear assignment* in the bipartite graph; a linear assignment is a subset of edges such that each node in the graph have exactly one edge assigned. Fig. 5 shows an example of the linear assignment corresponding to the coding template sequence shown in Fig. 2.

Each edge in the bipartite graph is assigned a weight which corresponds to a cost in terms of rate and distortion of the quantization (direct or differential) or component rejection represented by the edge. Still focusing on the middle subgraph of Fig. 4, edges representing differential encoding possibilites have weights of the type:

$$w = d^{i,m}_{j,m+1} + \lambda r^{i,m}_{j,m+1}, \tag{4}$$

where $d^{i,m}_{j,m+1}$ denote the quantization distortion, and $r^{i,m}_{j,m+1}$ the number of bits needed for representing component j in segment $m+1$ differentially relative to component i in segment m. In a similar manner, edges representing direct encoding have weights of the form

$$w = d^{dir}_{j,m+1} + \lambda r^{dir}_{j,m+1}. \tag{5}$$

Edges representing discarding components have the following type of weights

$$w = d^{mod}_{j,m+1}, \tag{6}$$

where $d^{mod}_{j,m+1}$ represents the modeling distortion occuring by not including candidate j in the sinusoidal representation. Finally, edges which are not connected to the K_{m+1} candidate nodes $s_{1,m+1} \cdots s_{K_{m+1},m+1}$ all have weights $w = 0$.

Assuming that per segment distortions and rates can be found by adding up the distortions and rates related to each candidate component, it turns out that the problem in Eq. (3) is solved by finding in the bipartite graph of Fig. 5 the linear assignment with minimum total weight. This problem, which is often refered to as the linear assignment problem, is well-known within graph theory, and several schemes exist for solving it; assuming that each segment in super-segment l contains K_l candidate sinusoids, a solution can be found in $\mathcal{O}((3M_lK_l)^3)$ arithmetic operations, see e.g. [7].

In summary, the scheme for solving our original problem in Eq. (1) involves, for a given value of the Lagrange multiplier λ, the solution of L independent subproblems of the form in Eq. (3). Each of these subproblems can be formulated and solved as instances of the linear assignment problem. The solution of an assignment problem describes unambiguously the coding template sequence to be used in order to minimize the distortion under the rate constraint specified by the selected value of λ. Since, usually, this λ value does not correspond to the target rate of interest, λ may be updated using e.g. a bisection method, and the entire process repeated until the target rate of interest is reached. For more details on the proposed algorithm, the reader is refered to [8].

3. SIMULATION EXPERIMENTS

We evaluate the presented algorithm in simulation experiments with audio signal fragments, sampled at 44.1 kHz and with a duration of 7-13 seconds. The following signal fragments were included in the experiments: Abba, Celine Dion, German male speech, Metallica, and Suzanne Vega.

In all experiments, the input signal is segmented into non-overlapping fixed-length super-segments of 8192 samples (corresponding to 186 ms at a sampling rate of 44.1 kHz), which in turn are segmented into fixed-length segments of 1024 samples with an overlap of 50%, resulting in $M_l = 16$ segments per super-segment. For each segment the candidate component set is created by estimating the 80 perceptually most relevant sinusoidal components using the psycho-acoustic based matching pursuit algorithm described in [9].

In all experiments, amplitude parameters are quantized using a log-quantizer with a relative spacing between reconstruction points

System B	26.0	18.0	10.0
Abba, System A	25.1	17.3	9.5
Celine, System A	25.1	17.3	9.5
Speech, System A	25.3	17.4	9.6
Metallica, System A	25.4	17.5	9.6
Vega, System A	25.1	17.4	9.5

Table 1: Rates [kbps] needed with System A (the proposed algorithm) and System B for achieving similar distortion levels for different test signals.

of 1.6% (both for direct and time-differential quantization), while a log-quantizer with a relative output level spacing of 0.3% is used for frequency parameters. Direct and time-differential phase parameters are quantized using a 4 bit uniform quantizer. With these fixed quantizer settings, the quantized signals are perceptually identical to signals constructed from unquantized sinusoidal parameters.

With these settings, the proposed algorithm is run for a given target bit budget, and the selected sinusoids are TD-quantized and encoded according to the optimal coding template sequence found for each super-segment. Let us for later reference denote this system as 'System A'.

We compare System A with a variant of the TD sinusoidal coding system in [6]. This system has a potential weakness related to the way in which sinusoids are selected and distributed across segments. Specifically, in [6] sinusoids are distributed across the fixed-length segments according to perceptual relevance; the bit cost associated with individual sinusoidal components is *not* taken into account (in contrast to the presented algorithm). For later reference, we refer to this TD system as 'System B'.

The algorithms for determining TD-relations in Systems A and B both rely on edge weights reflecting the number of bits for encoding of direct and differential model parameters (Eqs. (5)–(6)) and distortion terms (Eqs. (4)–(6)). In practice, the rate values are found from look-up in pre-calculated Huffman code word tables, while all distortion terms are computed using the perceptual distortion measure described in [10].

For a given test signal and a given value of R_t, the proposed algorithm was executed, and the resulting total signal distortion (the distortion term in Eq. (1)) was noted. This procedure was repeated for a large range of target rates R_t, and for all test signals. Subsequently, distortion-rate (D-R) curves for each signal could be plotted by linearly interpolation between the obtained distortion-rate pairs. A similar procedure was followed for System B, resulting in another set of D-R curves. A general first observation related to these D-R curves is that the distortion values obtained with System A (the proposed algorithm) were always lower than those of System B, for any target bit rate; conversely, the bit rate needed for reaching a certain distortion level was always lower with the proposed algorithm.

Table 1 summarizes the information of the D-R curves by showing the bit rates needed with System A for achieving distortion levels corresponding to target rates of 26, 18, and 10 kbps with System B. For example, we see that when System B requires 26.0 kbps for encoding the Suzanne Vega signal at a certain distortion level, System A needs 25.1 kbps. In comparing the performance of Systems A and B we see from Table 1 that the proposed algorithm typically achieves a bit rate reduction of approximately 3% at bit rates around 26 kbps and roughly 5% around 10 kbps.

We can explain the larger efficiency of the proposed system towards lower bit rates by considering quasi-periodic signals, e.g. voiced regions of voice signals. For larger bit-rates, say 26 kbps, the bit budget allows for extraction of most of the harmonics in the signal, and the extracted sinusoids will form unbroken harmonic 'tracks' when plotted in the time-frequency plane. In this case, both Systems A and B finds that TD-encoding of sinusoids within each track is the most efficient. For lower rates, however, the bit budget does not allow for extraction of all harmonics. In System B, where sinusoids are extracted according to perceptual relevance only, the sinusoids do not necessarily form unbroken harmonic tracks, since the perceptually most important sinusoids do not necessarily belong to the same harmonic, when seen across a certain time duration. In this case, System B becomes less efficient, because TD-parameter differences will increase and thus (typically) become more bit rate expensive. System A, on the other hand, takes the associated bit cost into account when extracting sinusoids. As a consequence, the sinusoids chosen with this algorithm tend to form unbroken tracks (more tracks will occur for increasing bit rate), and the encoding of the sinusoids becomes more efficient.

While the problem formulation presented here assumed fixed quantizers for each sinusoid, it appears possible to generalize the algorithm such that individual sinusoidal quantizers are adapted according to the specified target bit rate. This remains a topic for further research.

4. CONCLUSION

In this paper we have presented an algorithm for jointly extracting and time-differential encoding of sinusoidal model parameters. For a pre-specified target bit rate, the algorithm selects from a set of candidate sinusoids for a number of consecutive segments the subset which leads to minimum distortion in the reconstructed signal. Simulation experiments show that the proposed algorithm leads to a reduction of 3-5% in bit rate compared to a state-of-the-art time-differential sinusoidal encoding system.

REFERENCES

[1] S. C. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, Ph.D. thesis, Stanford University, 1998.

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

[3] 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.

[4] 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.

[5] J. Jensen and R. Heusdens, "Schemes for optimal frequency-differential encoding of sinusoidal model parameters," *Signal Processing*, vol. 83, no. 8, pp. 1721–1735, August 2003.

[6] J. Jensen and R. Heusdens, "A comparison of differential schemes for low-rate sinusoidal audio coding," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[7] R. Jonker and A. Volgenant, "A shortest path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, pp. 325 – 340, 1987.

[8] J. Jensen and R. Heusdens, "Time-differential sinusoidal encoding of multiple consecutive signal segments," Tech. Rep. TR-2004-01, Technical University of Delft, 2004.

[9] 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.

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