SINUSOIDAL CODING USING LOUDNESS-BASED COMPONENT SELECTION

Heiko Purnhagen, Nikolaus Meine, Bernd Edler

Laboratorium für Informationstechnologie
University of Hannover
Schneiderberg 32, 30167 Hannover, Germany
{purnhage,meine,edler}@tnt.uni-hannover.de

ABSTRACT

Sinusoidal modelling forms the base of parametric audio coding systems, like MPEG-4 HILN, where it is combined with noise and transient models. A parametric encoder decomposes the audio signal into components that are described by appropriate models and represented by model parameters. To achieve efficient coding at very low bitrates, selection of the perceptually most relevant signal components (e.g. sinusoids) is essential, as only a limited number of component parameters can be conveyed in the bitstream. Various strategies for sinusoidal component selection have been proposed in the literature. This paper introduces a new, loudness-based strategy and tries to compare the different strategies using objective and subjective criteria.

1. INTRODUCTION

Parametric representations of speech and audio have long been used for signal analysis/modification/synthesis and to achieve efficient coding. In general, parametric representations are based on a decomposition of the input signal into components that are described by appropriate source models and represented by model parameters. Sinusoidal modelling, i.e. a parametric representation that utilises only sinusoidal components [1], is very popular because probably most real-world audio signals are dominated by tonal signal components. Equation 1 shows how the input signal $x(t)$ is approximated by a set of N sinusoids i with slowly varying parameters for amplitude $a_i(t)$ and frequency $f_i(t)$ and a start phase φ_i.

$$\hat{x}(t) = \sum_{i=1}^{N} a_i(t) \cdot \sin(\varphi_i + 2\pi\textstyle\int_0^t f_i(\tau)\,d\tau) \qquad (1)$$

However, noise-like and transient signal components can not be efficiently represented by sinusoidal modelling. Hence, a sinusoidal model is often combined with additional signal models for noise and transients. Overviews of such hybrid models can be found e.g. in [2, 3].

For the experiments reported in this paper, the MPEG-4 parametric audio coder HILN ("Harmonic and Individual Lines plus Noise") was used as a framework [4, 5].

In the HILN encoder, the input signal is decomposed into different signal components and then the model parameters for the components are estimated: *Individual sinusoids* are described by their frequencies and amplitudes, a *harmonic tone* is described by its fundamental frequency, amplitude, and the spectral envelope of its partials, and a *noise* signal is described by its amplitude and spectral envelope. The modelling of transient components is improved by optional parameters describing their temporal amplitude envelope. Finally, the component parameters are quantised, coded, and multiplexed to form a bitstream. The target bitrate range of HILN is approximately 6 to 16 kbit/s, and typically an audio bandwidth of 8 kHz and a frame length (hop size) of 32 ms are used.

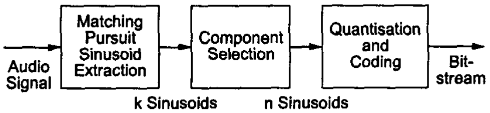

Fig. 1. General block diagram of encoder framework.

In the HILN decoder, the parameters of the components are decoded and then the component signals are re-synthesised according to the transmitted parameters. By combining these signals, the output signal of the HILN decoder is obtained. Sinusoidal components continued from the previous frame (i.e. that are part of a longer trajectory) are synthesised using interpolation of frequency and amplitude parameters to avoid phase discontinuities. For new ("born") sinusoids, usually the start phase parameters are not transmitted and random values are used instead.

This paper is structured as follows: Section 2 introduces the experimental setup and describes the different strategies for sinusoidal component selection considered here, including a new, loudness-based strategy. In Section 3, the performance of these strategies is compared using objective and subjective criteria. The paper ends with an outlook and conclusions in Section 4.

2. COMPONENT SELECTION STRATEGIES

This paper focuses on the problem that only a limited number of sinusoids can be transmitted in very low bitrate coding applications. To compare different sinusoidal component selection strategies required to address this problem, the fast HILN encoder presented in [6] was used as framework. For this purpose, the noise and harmonic tone components as well as the optional temporal envelopes for transients were disabled. The encoder uses "Matching Pursuit"-based extraction of sinusoidal components [7], followed by a psychoacoustic component selection and finally the quantisation and coding block, as outlined in Figure 1.

If only n out of k extracted sinusoids can be transmitted, the ideal component selection would be the one with smallest impact on the perceived sound. The following subsections describe different strategies for this component selection.

0-7803-7402-9/02/$17.00 ©2002 IEEE

2.1. Strategy SNR

The Matching Pursuit sinusoid extraction used in the first block of Figure 1 is a greedy algorithm that iteratively extracts sinusoids from the current frame of the input signal in order to minimise the energy of the residual, i.e. maximise the SNR (signal-to-noise ratio) of the approximation. Hence, sinusoids are extracted in the order of decreasing amplitudes, i.e. the sinusoids with the highest amplitude is extracted first. To reduce computational complexity, the Matching Pursuit is implemented in the frequency domain [6]. A sampling rate of 16 kHz, a window length of 64 ms and a hop size of 32 ms are used for all experiments. Since the k extracted sinusoids are already ordered according to the SNR strategy, the selection of n sinusoids is simply done by taking the first n entries of the ordered list of all k extracted sinusoids.

2.2. Strategy SMR

For this strategy [8], first the masked threshold $M_k(f)$ describing the simultaneous masking [11] caused by all k extracted sinusoids is calculated using a parametric psychoacoustic model. This approach allows more accurate modelling of $M_k(f)$ than possible with an FFT-based psychoacoustic model as known from MPEG-1/2 audio coding. The sinusoids are then re-ordered according to their SMR (signal-to-mask ratio) so that the sinusoid i with maximum $a_i/M_k(f_i)$ is selected first, etc.

2.3. Strategy HILN

During the development of HILN, a different masking-based strategy was used in the encoder [9]. It is an iterative algorithm where in the j-th step the sinusoid with maximum $a_i/M_{j-1}(f_i)$ is selected, i.e. the one which is highest above the masked threshold $M_{j-1}(f)$ caused by the $j-1$ sinusoids that were already selected in the previous steps. The iteration is started with the threshold in quiet $M_0(f)$. This can be considered as an algorithm to re-order the list of extracted sinusoids. Please note that, although this strategy is denoted as HILN here, an HILN encoder could of course employ any of the selection strategies discussed here.

2.4. Strategy ESW

This strategy was introduced in [10] and is named Excitation Similarity Weighting (ESW). It tries to maximise the matching between the auditory excitation pattern [11] associated with the original signal and the auditory excitation pattern associated with the selected sinusoids. For the experiments reported here, the set of all k extracted sinusoids was regarded as the original signal. To measure the similarity of the excitation patterns, the difference between the excitation levels in dB of the original and the selected sinusoids is accumulated along the basilar membrane. In each step of this iterative procedure, the sinusoid is selected which results in the best improvement in similarity. The procedure is equivalent to an iterative maximisation of the overall excitation level Q_{ESW} in dB

$$Q_{\mathrm{ESW}} = \int_0^{24\text{ Bark}} L_{\mathrm{E}}(z)\,dz \qquad (2)$$

where $L_{\mathrm{E}}(z)$ is the excitation level in dB at critical-band rate z.

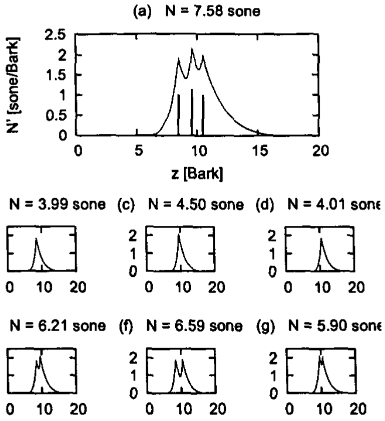

Fig. 2. Loudness N and specific loudness $N'(z)$ for combinations of 3 sinusoids (f = 1000/1250/1500 Hz, L = 60/62/60 dB).

2.5. Strategy LOUD

Inspired by the ESW strategy, the new selection strategy LOUD is proposed here, which tries to improve perceptual similarity even more. It uses the specific loudness $N'(z)$ in sone/Bark instead of excitation level $L_{\mathrm{E}}(z)$ in dB. Both $N'(z)$ and $L_{\mathrm{E}}(z)$ are non-linear functions of the excitation $E(z)$. Strategy LOUD results in a selection procedure that iteratively maximises the loudness N in sone

$$N = \int_0^{24\text{ Bark}} N'(z)\,dz \qquad (3)$$

that is associated with the n selected sinusoids.

2.6. Strategy LREV

All selection strategies discussed until now make use of greedy re-ordering algorithms that start with the selection of the most relevant component. However, since the masking- and excitation-based selection strategies utilise non-linear and non-orthogonal quality measures, the greedy approach can lead to sub-optimal results. This can be illustrated using the example shown in Figure 2, where k=3 sinusoidal components at 1000 Hz, 1250 Hz, and 1500 Hz are considered (a). Both the loudness-based strategy LOUD and subjective assessment indicate that choosing the sinusoids at 1000 Hz and 1500 Hz (f) is the optimum selection for n=2. All greedy re-ordering algorithms presented here, however, would select the sinusoid at 1250 Hz as first component, leaving only the sub-optimal alternatives (e) and (g) for n=2. One approach to this problem is to reverse the "direction" of the re-ordering procedures by starting from the full set of k sinusoids and iteratively de-selecting those components that considered of lowest relevance. Strategy LREV uses this reversed selection procedure applied to the loudness measure N.

2.7. Strategy LOPT

It should be obvious that also LREV is a greedy algorithm. To assess the sub-optimality of both strategies LOUD and LREV, a full search to find the best subset of n sinusoids that gives the highest loudness N was implemented as reference and is referred to as

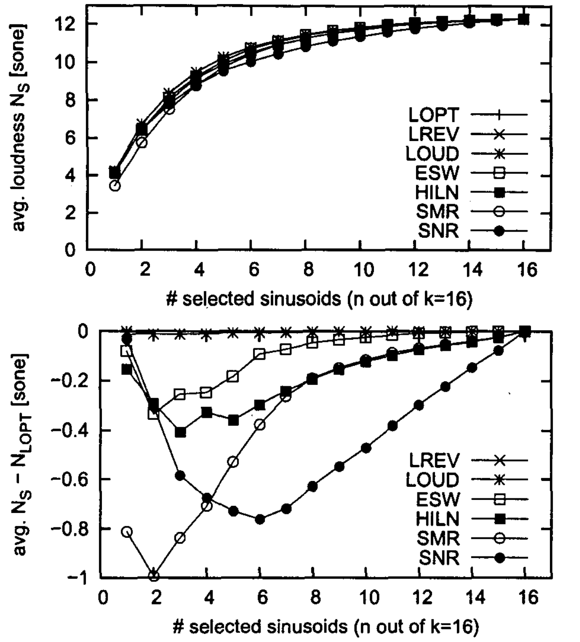

Fig. 3. Average loudness $\overline{N_S}$ and loudness difference $\overline{N_S - N_{\mathrm{LOPT}}}$ achieved by different strategies S for selection of n out of k=16 sinusoids.

strategy LOPT here. However, the computational complexity of the this search is $O(2^k)$, which becomes prohibitive for values of k above approximately 20. In addition, LOPT does not lead to a simple re-ordering of the list of sinusoids, as indicated in the example in Figure 2. Hence it can not be easily combined with the bit allocation strategies typically employed in the quantisation and coding block in Figure 1, as these often iteratively choose n such that a given bit budget per frame is not exceeded.

3. RESULTS

To compare the performance of the selection strategies described in Section 2, they were applied to a set of 12 speech and music items (total duration 141 s) which was also used throughout the MPEG-4 Audio core experiment procedure. Due to the lack of a simple, well-established perceptual similarity measure, an objective comparison is not easy. However, since most systems for the objective measurement of perceived audio quality, like [12], internally utilise modelling of excitation patterns, the loudness measure described in Subsection 2.5 seems to be suitable for this purpose. Hence, the full search strategy LOPT is considered as reference here. To make the comparison computationally feasible, the selection of n=1..16 sinusoids out of k=16 sinusoids extracted by the Matching Pursuit was assessed. Figure 3 shows the average loudness $\overline{N_S}$ and loudness difference $\overline{N_S - N_{\mathrm{LOPT}}}$ achieved by different strategies S. Table 1 gives the numerical values for n=8, including the maximum value of the loudness difference $N_S - N_{\mathrm{LOPT}}$ for all frames of the items.

It can be seen from Figure 3 that LOUD and LREV perform almost equal to LOPT, with a slight advantage for LREV. This means that extremely complex full search of LOPT gives only very little additional benefit. As expected, ESW behaves quite similar to LOUD; only for small values of n, differences in the achieved loudness N are observed. The masking-based strategies SMR and HILN perform almost identical for n=7..16, but not as good as the excitation-based strategies LOUD and ESW. It is interesting to observe that SMR shows the worst performance of all strategies for n=1..4. Strategy SNR shows the worst performance of all strategies for n5..16. Please note that the vanishing differences in performance when n reaches k=16 are inherently caused by the experimental setup, i.e. the n out of k selection.

Strategy S	avg. loudness $\overline{N_S}$	avg. diff. $\overline{N_S - N_{\mathrm{LOPT}}}$	max diff. $N_S - N_{\mathrm{LOPT}}$
SNR	10.833 sone	-0.628 sone	-5.979 sone
SMR	11.269 sone	-0.192 sone	-4.593 sone
HILN	11.267 sone	-0.194 sone	-4.006 sone
ESW	11.415 sone	-0.046 sone	-0.925 sone
LOUD	11.459 sone	-0.003 sone	-0.395 sone
LREV	11.460 sone	-0.001 sone	-0.237 sone
LOPT	11.461 sone	0.000 sone	0.000 sone
all 16	12.303 sone	0.842 sone	5.570 sone

Table 1. Average loudness $\overline{N_S}$ achieved by different strategies S for selection of n=8 out of k=16 sinusoids.

To illustrate the differences between the selection strategies discussed here, Figure 4 show the selected n=10 sinusoids out of k=40 extracted sinusoids for one frame of a pop music item with vocals. The re-ordered ranking is indicated by the labels 1..10 and the original spectrum as well as the masked threshold caused by the selected sinusoids are included in the graphs.

To allow subjective assessment of the different selection strategies, they were implemented in the HILN coder framework. Various test items were encoded at a bitrate of 6 kbit/s, i.e. using about 10 to 20 sinusoids per frame, and coding of the noise component was enabled. An informal listening test indicates that the strategy SNR results in a lower quality than all other strategies. However, the differences in subjective quality between the other strategies are fairly subtle and probably would show no statistically significant grading differences in a formal listening test.

4. CONCLUSIONS

In this paper, a strategy for component selection in a sinusoidal coder was introduced that tries to find the subset of the extracted sinusoids that is perceptually most similar to complete set of extracted sinusoids by approximating the auditory excitation pattern, using loudness difference as similarity measure. It was found that a full search for the optimal subset provides almost no improvement over the greedy re-ordering algorithms normally used. The loudness-based strategy is similar to another excitation-similarity-based strategy (ESW) introduced in [10]. While competitive otherwise, masking-based strategies like SMR achieve lower performance according to the loudness difference measure for a low number of selected sinusoids. Strategy SNR, which minimises the mean square error of the residual, achieves lowest performance according to both the loudness difference measure and subjective

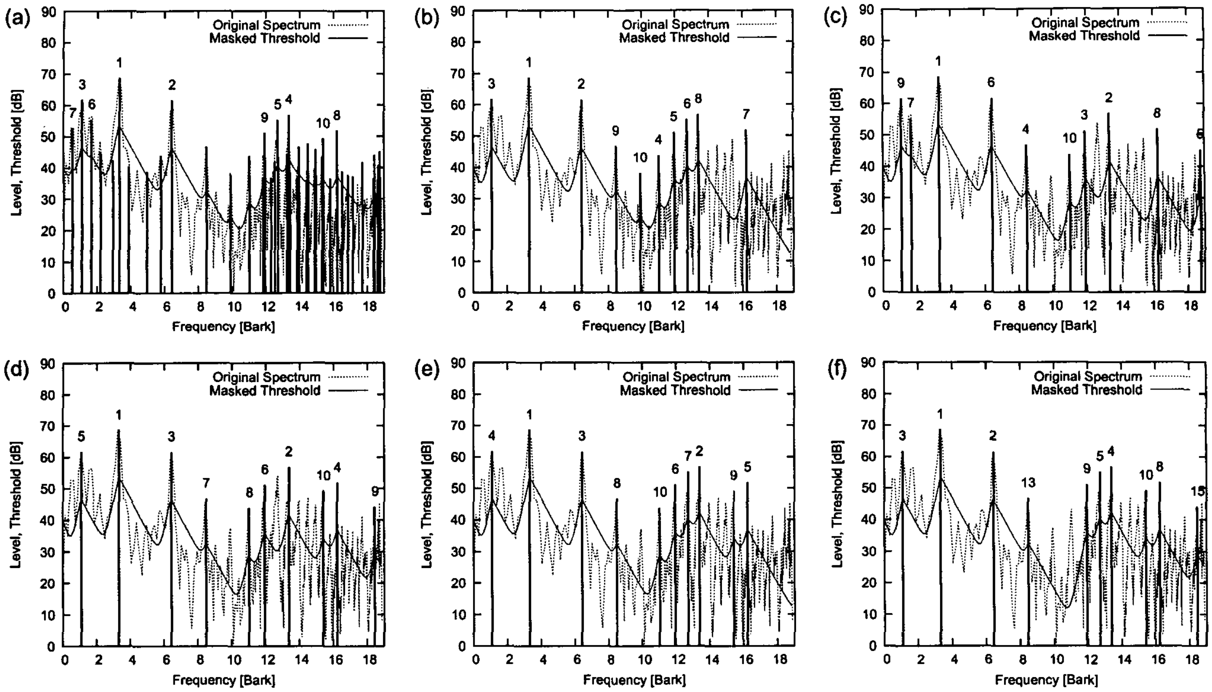

Fig. 4. (a) 40 sinusoids selected by strategy SNR (i.e. original ordering from Matching Pursuit, first 10 sinusoids labelled); (b) 10 out of 40 sinusoids selected by SMR; (c) 10 out of 40 sinusoids selected HILN; (d) 10 out of 40 sinusoids selected ESW; (e) 10 out of 40 sinusoids selected by LOUD and LREV; (f) 10 out of 20 sinusoids selected by LOPT (labels show rank in original ordering from Matching Pursuit).

assessment. This indicates that loudness-based similarity measure can be appropriate for the applications considered here.

In the experiments reported in this paper, the component selection was carried out for each frame independently. For a sinusoid that is part of a trajectory continuing over several frames, this can lead to an annoying unsteadiness if it is selected only intermittently. In [13], a strategy for selection of trajectories depending upon their duration and SMR was successfully utilised. Since models for the temporal effects of loudness summation are also known [11], the loudness-based selection strategy can be extended to rank the relevance of complete sinusoidal trajectories.

5. REFERENCES

[1] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. ASSP*, Vol. 34, No. 4, pp. 744–754, Aug. 1986.

[2] X. Rodet, "Musical Sound Signals Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models," *Proc. IEEE Time-Frequency and Time-Scale Workshop (TFTS'97)*, Coventry, Aug. 1997.

[3] H.. Purnhagen, "Advances in Parametric Audio Coding," *Proc. IEEE WASPAA*, Mohonk, Sep. 1999.

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

[5] H. Purnhagen and N. Meine, "HILN - The MPEG-4 Parametric Audio Coding Tools," *Proc. IEEE ISCAS 2000*, Geneva, May 2000.

[6] H. Purnhagen, N. Meine, and B. Edler, "Speeding up HILN – MPEG-4 Parametric Audio Encoding with Reduced Complexity," *AES 109th Convention*, Preprint 5177, Los Angeles, Sep. 2000.

[7] M. Goodwin, *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*, PhD thesis, University of California, Berkeley, 1997.

[8] T. Verma and T. Meng, "Sinusoidal Modeling using Frame-Based Perceptually Weighted Matching Pursuits," *Proc. IEEE ICASSP*, Phoenix, Mar. 1999.

[9] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC - Analysis/Synthesis Audio Codec for Very Low Bit Rates," *AES 100th Convention*, Preprint 4179, Copenhagen, May 1996.

[10] T. Painter and A. Spanias, "Perceptual Segmentation and Component Slection in Compact Sinusoidal Representations of Audio," *Proc. IEEE ICASSP 2001*, Salt Lake City, May 2001.

[11] E. Zwicker and H. Fastl, *Psychoacoustics – Facts and Models*, 2nd Ed., Springer, Berlin, 1999.

[12] T. Thiede et al., "PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality," *Journal AES*, Vol. 48, No. 1/2, Jan./Feb. 2000.

[13] S. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, PhD thesis, CCRMA, Stanford University, Dec. 1998.