PARAMETRIC AUDIO CODING

Bernd Edler, Heiko Purnhagen

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

ABSTRACT

For very low bit rate audio coding applications in mobile communications or on the internet, parametric audio coding has evolved as a technique complementing the more traditional approaches. These are transform codecs originally designed for achieving CD-like quality on one hand, and specialized speech codecs on the other hand. Both of these techniques usually represent the audio signal waveform in a way such that the decoder output signal gives an approximation of the encoder input signal, while taking into account perceptual criteria. Compared to this approach, in parametric audio coding the models of the signal source and of human perception are extended. The source model is now based on the assumption that the audio signal is the sum of "components," each of which can be approximated by a relatively simple signal model with a small number of parameters. The perception model is based on the assumption that the sound of the decoder output signal should be as similar as possible to that of the encoder input signal. Therefore, the approximation of waveforms is no longer necessary. This approach can lead to a very efficient representation. However, a suitable set of models for signal components, a good decomposition, and a good parameter estimation are all vital for achieving maximum audio quality.

We will give an overview on the current status of parametric audio coding developments and demonstrate advantages and challenges of this approach. Finally, we will indicate possible directions of further improvements.

1. INTRODUCTION

For high quality coding of arbitrary audio signals transform coding is widely used [1], since it allows efficient reduction of redundancy and irrelevancy based on the spectral decomposition of the audio signal. However, if the target bitrate is reduced to about 16 kbit/s or below, this technique is no longer optimal for all types of audio material. As an alternative, speech coders are frequently used in this bitrate range, but due to the specialized structure their efficiency highly depends on the characteristics of the input signal. Currently there is no audio coding technique with sufficient generality for the very low bitrate range, and even in combination the two mentioned techniques do not completely cover all types of audio material.

As a consequence the interest in parametric audio coding as a third technique has grown during the last years. While first implementations mainly focused on speech coding [2], it is meanwhile extended towards applicability for arbitrary audio signals [3]. With the so-called HILN (Harmonic and Individual Lines plus Noise) tools [4], a parametric coding scheme recently was accepted for inclusion in Version 2 of the MPEG-4 Audio Standard [5]. By combining all three coding techniques MPEG-4 is able to provide efficient representations for all types of audio material at bitrates down to 6 kbit/s.

This paper first shows the basic principles of parametric audio coding and gives a comparison with the other two coding techniques mentioned above (Section 2). This is followed by an overview of specific source models and perception models used in various approaches to parametric coding (Section 3). Section 4 gives more details of the MPEG-4 HILN tools. The achieved coding efficiency and its dependency on the audio signal content is addressed in Section 5. Section 6 concludes with a summary of the current status and thoughts on possible future developments.

2. FUNDAMENTALS OF PARAMETRIC AUDIO CODING

Bitrate reduction in audio coding systems usually is based on the concepts of redundancy reduction and irrelevancy reduction. Redundancy reduction is achieved by exploiting characteristics of the input signal. Based on a model of the signal source it reduces the bitrate without any loss of accuracy. For example, the basic structure of CELP (Code Excited Linear Predictive) speech coders is based on a model of the human vocal tract [6]. Different types of excitation are used for periodic (voiced) components and for non-periodic (unvoiced and plosive) components. The variability of resonances is resembled by the adaptive predictor structure. Transform coders designed for arbitrary audio signals cannot use such an explicit source model due to the fact that different musical instruments can have totally different characteristics. Therefore the more generic assumption is made that the signal is quasi-stationary. This means that the signal characteristics stay nearly constant within short time intervals. Then a coding gain can be achieved for signals with non-flat short term spectra by a spectral decomposition and appropriate coding. Most coders which use a high spectral resolution (e.g. > 500 lines) can switch to a reduced spectral resolution in order to increase the temporal resolution in the presence of strong transients in the input signal. On the other hand, if the signal characteristics remain constant for several subsequent intervals, additional redundancy reduction is possible, e.g. using predictive techniques.

Irrelevancy reduction usually is achieved by a controlled reduction of the accuracy, which takes into account properties of the human auditory system. The goal is to keep distortions inaudible or to let them sound as pleasant as possible. For this purpose a model of human sound perception is needed. Speech coders usu-

ally are based on a relatively simple model which assumes that distortions should have a temporal envelope and spectral shape similar to that of the input signal. Transform coders usually have built in a more sophisticated perception model which individually controls the quantization of the spectral components. For this purpose a signal dependent spectrally and temporally varying masking threshold is estimated in the encoder indicating the amount of just not yet audible distortion.

Compared to the relatively inflexible spectral decomposition using a transform, the idea of parametric audio coding is to decompose the audio signal in a more adaptive way. For this purpose different source models for single components are developed and parameters are defined which allow to describe the actual characteristics. After the decomposition the parameters have to be quantized, encoded, and transmitted. The most obvious decomposition might seem to be a separation into single instrument signals. However there would be several difficulties in such an approach. First of all, there are currently no separation algorithms which are reliable enough to cope with complex sound mixtures, although research in the field of auditory scene analysis is dealing with this problem. Furthermore the exact description of the sounds of the individual instruments would require a high number of parameters and thus seems not to be suitable for an efficient representation. Therefore more general signal models for single components are used, like the examples described in Section 3. A parametric audio decoder has to provide synthesizers for the different component types, which are controlled by the decoded parameters.

The different signal representation in parametric audio coding requires new approaches for the consideration of perceptual criteria. One quite obvious approach is to consider the influence of parameter deviations on the perception of the synthesized sounds in the design of the quantizers. Another very important issue however is the selection of components for which parameters are to be transmitted to achieve the optimum subjective quality at a given bitrate. For this purpose a perceptual model for the relevancy of signal components is required.

3. SOURCE AND PERCEPTION MODELS FOR PARAMETRIC AUDIO CODING

Parametric audio coding heavily relies on the availability of source models which allow the description of signal components with a small number of parameters. In the following a brief overview of applicable models is given, while more details can be found in [7]. In addition specific perception models are described which are needed for efficiency with respect to subjective quality.

3.1. Physical Models for Excitation and Resonances

The sound generation of many musical instruments can be described by a single pulse-like or a periodic excitation in conjunction with a resonance body or a combination of multiple resonances. An example of such a system is a plucked string instrument where the plucking is a pulse-like excitation and the resonances of the string and the body shape the sound.

However this approach faces problems in the separation into real instruments, in the classification of the instruments, and in the parameter estimation. Therefore its application in coding of natural signals is currently restricted to speech coders modeling the excitation and the vocal tract as described above.

3.2. Sinusoidal Models

An alternative to the separation into single instrument signals is the decomposition into signals which can be described with relatively simple mathematical models. An approach for a tonal signal $x(t)$ is to regard it as a superposition of N individual sinusoidal components, each of which is described by slowly varying parameters for amplitude $a_i(t)$ and frequency $f_i(t)$ and a constant start phase φ_i:

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

This modeling approach originally used for musical instrument analysis/synthesis [8] was later applied in speech coding [2] and audio coding [3].

In coding applications, usually one set of model parameter values is transmitted per frame (i.e. time interval). In the decoder the parameters are interpolated between frames, if a sinusoid in one frame is continued from the previous frame. In this case only frequency and amplitude changes need to be transmitted. Multiresolution sinusoidal modeling uses frequency dependent frame lengths, so that high-frequency sinusoids are modeled with finer time resolution than low-frequency sinusoids.

3.3. Models for Transients

The time resolution usually selected for relatively stationary tonal components only provides a poor representation for transients, i.e. signal components with rapidly changing amplitudes. Therefore a special treatment for transients is required.

One approach is to use sinusoids in conjunction with an amplitude envelope function, which e.g. can be described by parameters for its temporal position and its attack and decay rates [3]. A second approach approximates the DCT spectrum of the transient using a sinusoidal model. Since the spectrum of a transient does not consist of distinctive lines, a third alternative is to use techniques known from transform coding to transmit the whole spectrum.

3.4. Noise Models

With respect to perception, a waveform approximation for noise signals is not necessary. Thus noise components can be modeled by a random signal with appropriate spectral and temporal envelopes.

The spectral noise shaping can be performed efficiently with filter structures as they are used in LPC (Linear Predictive Coding) based speech coders. The filter parameters also can be represented using techniques known from speech coding. The use of so-called reflection coefficients enables an easy adaptation of the filter order to the required level of spectral detail [4]. Some alternative methods are piece-wise linear noise spectrum approximation, Bark-band noise modeling [9], or DCT modeling of the noise spectrum [4].

3.5. Extended Sinusoidal Models

The model of individual sinusoids for tonal components has the advantage of a high flexibility in the presence of arbitrary mixtures of different instrument signals. However, the bitrate required for the transmission of their parameters increases nearly linearly with the number of sinusoids. On the other hand, most musical instruments produce harmonic tones consisting of partials at multiples of the

fundamental frequency f_0. Therefore it is useful to extend the sinusoidal model towards a harmonic model with parameters for f_0, and for the amplitudes a_i and phases φ_i of the partials. Replacing the amplitudes a_i by parameters representing the spectral envelope can increase the efficiency even further. Again LPC based models can be applied [4]. A possible alternative is derived from techniques known as "FM-synthesis" for music synthesizers [10].

The sinusoidal model also can be extended towards so-called damped sinusoids [9] which can improve the efficiency for non-stationary signals. In a further extension called "bandwidth-enhanced sinusoids" the frequencies and/or amplitudes of sinusoids can be modulated with a low-pass filtered noise [11] in order to allow a smoother transition between tonal and noise-like components.

3.6. Perception Models for Parameter Quantization

An efficient representation requires quantizers which are either designed taking into account perceptual criteria or even adaptively adjusted to the current signal content. For many of the parameters used in the models described above it is sufficient to design quantizers according to the sensitivity of the human ear in detecting deviations. Some of the resulting quantizer characteristics are summarized in the following.

All quantizers for frequency and amplitude parameters should be adjusted to the audibility thresholds for deviations known as "just noticeable differences." For frequencies the quantizer step sizes therefore should be approximately proportional to the frequency dependent critical band width. For amplitude parameter quantizers a logarithmic characteristic seems to be the most appropriate.

Subjective evaluations have shown that the relevancy of phase parameters of sinusoids generally is so low that they do not need to be transmitted. However in this case the temporal structure must be maintained by using an appropriate transient model and phase continuity must be guaranteed by frame-to-frame tracking.

As mentioned above, waveform approximation for noise components is not required. Thus only parameters for temporal and spectral envelopes need to be transmitted.

LPC parameters representing spectral envelopes can be encoded efficiently as so-called Logarithmic Area Ratios (LARs), which are based on a non-linear quantization of reflection coefficients. Alternatively a conversion to so-called Line Spectrum Frequency (LSF) parameters can be performed. While LSFs provide a slightly higher coding efficiency, LARs have advantages if differential encoding needs to be combined with a variable filter order [4].

3.7. Perception Models for Component Selection

As described above, the selection of components for which parameters are transmitted is a very important issue in parametric audio coding. For simplicity reasons this is illustrated for the representation of a signal segment by a relatively low number of individual sinusoids.

The first approach might be to select the components with the highest amplitude. For a signal with low-pass characteristic this procedure would obviously lead to a strong bandwidth reduction.

Another idea might be to apply a perception model to the input signal like it is done in transform coders. Now the selection could be based on the ratio of amplitudes and masked threshold. However this might have the effect that lines which are close to each other reduce each others relevancy so that none of them would be selected.

A way to overcome the described problems is to use a recursive procedure in which a perception model is applied only to components which already have been selected for transmission. In each step the one component is selected which has the maximum amplitude ratio over the masked threshold caused by all previously selected components [3].

4. HILN - PARAMETRIC AUDIO CODING IN MPEG-4

The MPEG-4 Audio Standard provides tools for coding of natural and synthetic audio objects and composition of such objects into an "audio scene" [5]. Natural audio objects such as speech and music can be coded at bitrates ranging from 2 kbit/s to 64 kbit/s and above using parametric speech coding (HVXC), CELP-based speech coding, parametric audio coding (HILN) or transform-based general audio coding (AAC, TwinVQ). The acronym HILN stands for "Harmonic and Individual Lines plus Noise." This name already gives an indication of the underlying source models, which are harmonic tones, individual sinusoids, and noise components.

All parameters are quantized according to perceptual criteria described in Section 3.6. The spectral envelopes of harmonic and noise components are represented by LARs for filters of variable order. Entropy coding of parameter changes for components present in subsequent frames further increases the efficiency. All new individual sinusoids are transmitted in an order with ascending frequencies to allow the use of an efficient technique called Sub-Division Coding [4].

Transients can be handled by an optional set of parameters describing the temporal envelope within a frame. If in one frame envelope parameters are present, additional flags are transmitted indicating the components to which the envelope has to be applied.

The block diagram of the HILN parametric audio decoder is shown in Figure 1. First the parameters of the components are decoded, then the component signals are synthesized and added to give the decoder output signal.

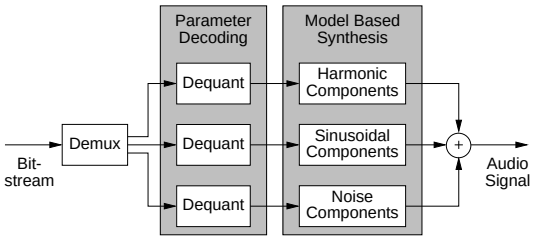

Figure 1: Block diagram of an HILN decoder.

In contrast to the decoder, the encoder is not fixed by the normative part of the standard. Here an encoding process is presented, where the first step is an analysis/synthesis loop for extracting individual sinusoids taking into account perceptual criteria as described in Section 3.7. In a second step a harmonic component is determined by searching for groups of individual sinusoids which could be partials with a common fundamental frequency. The residual signal after the extraction of tonal components is then regarded to be a noise component (Figure 2).

Besides pure bitrate reduction HILN provides additional functions like the possibility of an independent manipulation of speed and pitch during playback. Furthermore the bitstream is organized

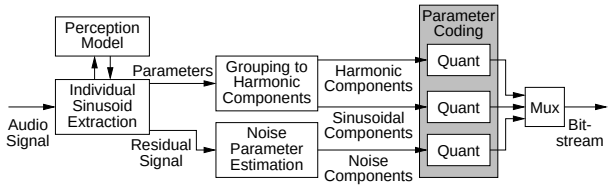

Figure 2: Block diagram of an HILN encoder.

in different sensitivity classes in order to allow unequal error protection. Bitrate scalability is supported by embedding a base layer and one ore more enhancement layers in a bitstream so that subsets at lower rates can be decoded.

5. CODING EFFICIENCY

The component selection procedure described above already gives a first indication for the expected coding efficiency of parametric audio coding in comparison to transform coding. An input signal only containing a relatively low number of significant components well matched by the models can be represented very efficiently. Transform coders however often need a relatively high overhead to transmit either all spectral components or appropriate side information. On the other hand parametric audio coding is less efficient for signals containing very complex sound mixtures or components not matched by the models.

This tendency is also reflected in the results of subjective evaluations [12] carried out during the MPEG-4 standardization process, in which HILN was compared to TwinVQ at 6 kbit/s (Figure 3). Additional informal comparisons showed the expected tendency that CELP at 6 kbit/s performs clearly better than the two audio coders for the speech item, but worse for most of the musical signals.

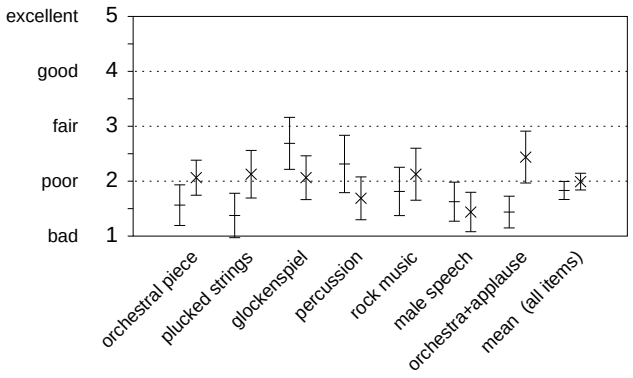

Figure 3: MPEG-4 Version 2 verification test results for HILN (–) and TwinVQ (x) at 6 kbit/s showing mean grades and 95% confidence intervals for 16 listeners (from [12]).

6. CONCLUSIONS

Although parametric audio coding is a relatively new technique it already has proven to be useful, especially at very low bitrates. The availability of various source models enables high flexibility and the integration of perception models helps to achieve a high efficiency with respect to subjective audio quality. With HILN a parametric audio coding scheme has been accepted as a part of the MPEG-4 Audio standard. Subjective evaluations have shown that HILN can complement speech and transform coders in a way that a combination of these three techniques is able to cover a wide variety of input signal types. However automatic codec selection techniques still need to be subject of further research. For the improvement of standalone parametric audio coders further investigations might focus on the parameter estimation for complex sound mixtures. Additionally an extension of the perception models towards a joint assessment of multiple frames could increase the "temporal stability" of the reconstructed sound.

REFERENCES

[1] K. Brandenburg and M. Bosi, "Overview of MPEG Audio: Current and Future Standards for Low Bit Rate Audio Coding," *J. Audio Eng. Soc.*, Vol. 45, No. 1/2, pp. 4–21, Jan./Feb. 1997.

[2] 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.

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

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

[5] R. Koenen, *Overview of the MPEG-4 Standard*, ISO/IEC JTC1/SC29/WG11 N3156, Dec. 1999. http://www.cselt.it/mpeg/standards/mpeg-4/mpeg-4.htm

[6] B. Edler, "Speech Coding in MPEG-4," *Int. J. of Speech Technology*, Vol. 2, No. 4, pp. 289–303, May 1999.

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

[8] J.-C. Risset and M. V. Matthews, "Analysis of musical-instrument tones," *Physics Today*, Vol. 22, pp. 22–30, Feb. 1969.

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

[10] B. Winduratna, "FM Analysis/Synthesis Based Audio Coding," *AES 104th Convention*, Preprint 4746, May 1998.

[11] K. Fitz and L. Haken, "Bandwidth Enhanced Sinusoidal Modeling in Lemur," *Proc. ICMC*, 1995.

[12] ISO/IEC, *Report on the MPEG-4 Audio Version 2 Verification Test*, ISO/IEC JTC1/SC29/WG11 N3075, Dec. 1999. http://www.tnt.uni-hannover.de/project/mpeg/audio/public/w3075.pdf

Further references and related links can be found at:
http://www.tnt.uni-hannover.de/~purnhage/