Progress on Parametric Coding for High Quality Audio

E.G.P. Schuijers[1], A.W.J. Oomen[1], A.C. den Brinker[2], D.J. Breebaart[2]
[1]*Philips Digital Systems Laboratories, Glaslaan 2 (SFJ7), 5616 LW Eindhoven, The Netherlands*
[2]*Philips Research Laboratories, Prof. Holstlaan 4 (WY82), 5656 AA Eindhoven, The Netherlands*
e-mail: erik.schuijers@philips.com

Abstract

In the scope of the "MPEG-4 Extension 2" standardisation process, a parametric coding scheme is under development. This coding scheme is based on the assumption that any audio signal can be dissected into three objects: transients, sinusoids and noise. For stereo signals, an additional modelling of the binaural cues is employed. Each of these objects allows for an efficient parametric representation. The parametric coder is targeting medium to high quality for CD quality material, at bit rates around 24 kbit/s stereo.

1. Introduction

In the quest for lower bit rates, the coding efficiency of traditional waveform coding schemes seems to be saturating. Parametric audio coding, however, promises a further increase in coding efficiency.

In the context of MPEG-4 Extension 2, Philips has proposed a parametric coding scheme that is based on the assumption that any audio signal can be decomposed into three objects: transients, sinusoids and noise. In addition, the model also employs an efficient parametric description of the stereo image. In this paper the parametric model will be described and two additional post-processing features, pitch and tempo scaling, will be presented. Furthermore, the current status in MPEG is discussed and subjective test results are presented.

2. Parametric Audio Model

The performance of a parametric audio coder is largely dependent on how well the model can represent the audio. Therefore the parameterised functions used to describe an audio signal have been chosen to reflect attributes that are well known from auditory perception and physics of natural audio signals. In summary, the attributes that are defined are transients, tonal components, noise-like sounds and spatial information. A short explanation of these attributes is given below; more details can be found in [1] and [2].

Transients; transients represent the non-stationary part of the audio signal. Transients are characterised by a fast change in signal spectrum. Modelling transients using quasi-stationary patterns proves to be an inefficient approach.

Transients can be roughly catagorised into two types. The first is characterised as a short burst of energy while the second can be described as a sudden change of signal level, i.e., a transition. According to this classification two types of transients have been defined by the coder: a "Meixner transient" [3] and a "step transient". The Meixner transient describes part of the waveform, the step transient only influences the way other objects (sinusoids and noise) are synthesised.

Sinusoids; sinusoids are used to represent tonal components that are clearly defined in frequency and typically last for a long time. Because of this clear definition in frequency, it is assumed that these spectral events can be modelled accurately using sinusoids.

For stationary segments, we use the following model for the deterministic part of the signal:

$$s(t) = \sum_{i=1}^{I(t)} A_i(t)\cos(\Phi_i(t)), \qquad \text{eq. 1}$$

with

$$\Phi_i(t) = \varphi_{s,i} + \int_{t_{s,i}}^{t} \omega_i(\tau)d\tau, \qquad \text{eq. 2}$$

where the subscript i denotes the i^{th} sinusoid, $A_i(t)$ represents the (slowly varying) amplitude, $\Phi_i(t)$ represents the phase function with start phase $\varphi_{s,i}$ and $\omega_i(t)$ represents the slowly varying frequency. $I(t)$ denotes the number of sinusoids at time t. The sinusoidal parameters are estimated on a frame by frame basis, which corresponds to sampling the functions $A_i(t)$, $\omega_i(t)$ and $\Phi_i(t)$ at a specific update rate.

Tracks that span multiple frames can be formed by a linking mechanism. The major part of the bit-rate savings stems from the fact that for these tracks only small changes need to be coded. Furthermore, if the assumption of slowly varying frequencies over time is true, the phase information becomes redundant.

Noise; noise represents the stochastic part of the audio signal. In nature, noise-like sources are often encountered, e.g. the rustle of the wind or unvoiced speech. The perception of such noise-like signals clearly differs from tonal signals.

In order to preserve or reproduce the perception of noise-like signals it is not necessary to precisely match the original waveform. It is sufficient to match only the spectral and temporal envelope. This makes the bit-rate requirements low.

In the parametric model, the spectral envelope is coded by means of a Laguerre model [4], the temporal envelope by an LPC model [5].

Stereo cues; these capture the perceived position and diffuseness of the spatial dimension contained in a stereo signal.

Instead of using (standardised) methods such as Mid/Side coding and intensity stereo to enable (spatial) information reduction, a parametric representation of the spatial information is used. This method aims at extending a mono audio stream, encoded as described above, with parameters describing the (perceptually relevant) spatial properties of the original stereo input material.

In this context, so-called binaural cue coding (BCC) schemes have been presented [6]. These schemes encode frequency and time-dependence of time and level differences between the input signals. Own research in the field of efficient stereo coding has demonstrated that some of the disadvantages of these binaural cue coding algorithms can be reduced by introducing a third signal parameter. This parameter aims at describing the difference between the two input signals that cannot be attributed to the encoded level and time (or phase) differences between the input channels. A suitable measure for this purpose is the inter-channel *coherence*. The three stereo parameters are analysed as a function of frequency and time. Subsequently, a mono signal is generated from the stereo input, which is encoded with the parametric encoding scheme as described above. The resulting parametric description of the mono signal is combined with the (quantized) spatial parameters to form the overall bit stream. The decoder basically comprises the reverse process: the incoming bit stream is split in mono and stereo parameters. First, the mono parameters are used to synthesize the mono signal. Subsequently, at the output the stereo signal is generated using the stereo parameters.

3. Post-processing

The parametric representation of audio as described in the previous sections is well suited for a number of post-processing operations on the audio signal. It is particularly suitable for time scaling and pitch scaling.

Time scaling; the goal of time scaling audio signals is to vary the duration of the signal, while providing the same perception of pitch as the original signal.

In order to enable time-scaling, the stereo and sinusoidal synthesis windows and noise temporal envelopes are scaled in the decoder with the appropriate (possibly time varying) factor. The time domain envelope of the transient object is not scaled. Perceptually this gives a more natural result. Furthermore, in order to prevent discontinuities in the signal, the phase of a sinusoid going from one frame to the next is adjusted to ensure a smooth transition from one frame to the next.

Pitch scaling; the goal of pitch scaling audio signals, is to vary the perceived pitch while keeping the duration of a signal constant.

Only the lower harmonics contribute to the percieved pitch according to [7]. These harmonics are modelled by the sinusoidal object; the noise object typically models the high-frequency range. Therefore it was decided not to change the noise parameters when using pitch scaling. A pitch change is thus implemented by scaling all the sinusoidal frequencies in a frame with the appropriate (possibly time varying) factor.

As time and pitch scaling are more or less ‘orthogonal’ procedures they can be combined to virtually any combination.

4. MPEG

In response to a Call for Proposals issued in January 2001 by MPEG, Philips has submitted a coding scheme on high-quality parametric audio coding. This submission was accepted as a new work-item for MPEG-4 Extension 2. In the mean time it has evolved into the parametric coding scheme as described in this paper. As of December 2002, MPEG-4 Extension 2 has been promoted to the more formal Committee Draft stage of an amendment. MPEG-4 Extension 2 is expected to reach the Final Draft International Standard (FDIS) stage by December 2003. A verification test will be conducted to prove its added value.

5. Test results

In order to assess the subjective quality of the parametric audio coder, an informal subjective listening test was conducted. In this test, parametric coding at 24 kbit/s stereo was tested together with MPEG-4 AAC at both 24 kbit/s stereo and 32 kbit/s stereo. France Telecom was kind enough to provide state-of-the-art encoded material using their MPEG-4 AAC encoder. In total 12 critical items, which are commonly used in the MPEG standardization process have been assessed by 10 subjects according to the MUSHRA testing methodology using the 100 points scale. The overall test results are presented in Figure 1. It clearly shows that parametric coding outperforms AAC at this bit rate.

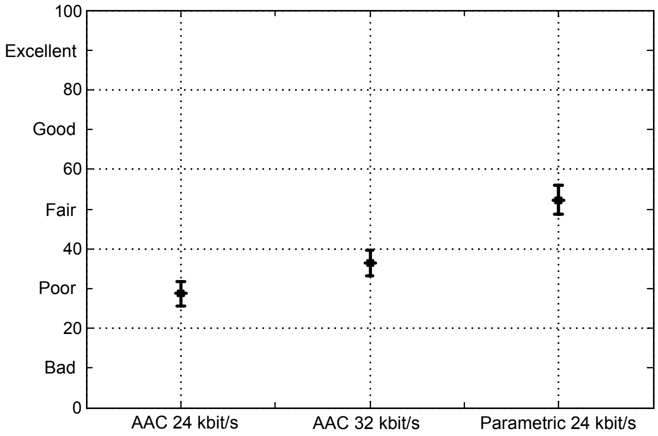

Figure 1: Subjective listening test results

6. Conclusions

Parametric audio coding is gaining interest and is showing clear progress in its performance. It is reaching an acceptable quality level for streaming applications. In its current status the parametric coding scheme at least outperforms MPEG-4 AAC at 24 kbit/s. Currently, the parametric coding scheme is under standardisation in the context of MPEG-4 Extension 2.

As an additional feature, parametric audio coding inherently allows for simple time and pitch scaling.

References

[1] E.G.P. Schuijers, A.W.J. Oomen, A.C. den Brinker and A.J. Gerrits, Advances in parametric coding for high-quality audio, *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, pp. 73–79, Leuven (B), November 15, 2002.

[2] E.G.P. Schuijers, A.W.J. Oomen, A.C. den Brinker and D.J. Breebaart, Advances in parametric coding for high-quality audio, *114th AES Convention*, Amsterdam (NL), March 22-25, 2003.

[3] A.C. den Brinker, Meixner-like functions having a rational z-transform, *Int. J. Circuit Theory Appl.*, 23:237–246, 1995.

[4] V. Voitishchuk, A.C. den Brinker and S.J.L. van Eijndhoven, Alternatives for warped linear predictors, *Proc. 12th ProRISC Workshop on Circuits, Systems and Signal Processing*, pp. 710–713, Veldhoven (NL), November 29-30, 2001.

[5] J. Herre and J.D. Johnston, Enhancing the performance of perceptual audio coders by using temporal noise shaping, *Preprint 4384, 101st AES Convention*, Los Angeles (USA), 8-11 November 1996.

[6] C. Faller and F. Baumgarte, Efficient representation of spatial audio using perceptual parameterization, *WASPAA, workshop on applications of signal processing on audio and acoustics*, New Paltz, New York, October 21-24, 2001.

[7] J.L. Goldstein An optimum processor for central formation of pitch of complex tones, *J. Acoust. Soc. Am.*, 46:1496–1516, 1973.