Very Low Bit Rate Audio Coding Development

Bernd Edler

Laboratorium für Informationstechnologie

University of Hannover, Germany

Abstract

Audio coding at very low bit rates requires very efficient evaluation of characteristics of the signal sources and of perceptual properties of the human ear. An overview of models for the signal source employed in different coding schemes is given. Some of the basic concepts for perceptual quantization and coding of model parameters are shown, including approaches for scaleability with respect to bit rate and quality.

1 Introduction

For internet and multimedia applications new demands for audio coding arise. On one hand a very efficient representation is required in order to cope with limited bandwidth. This includes coding of music signals at bit rates which up to now were mainly used for the transmission of speech. On the other hand high interactivity is desirable, which allows flexible access to coded data and manipulation by the recipient. Currently audio codecs based on various different approaches are available. This paper tries to give an overview on the underlying techniques and to give an outlook on possible future improvements.

The process of bit rate reduction in low and very low bit rate audio coding is based on the following two methods known from information theory: The first method, the so-called redundancy reduction exploits characteristics of the input signal and allows to represent the signal with reduced bit rate without introducing any degradation of the signal. This can be achieved by transforming the PCM samples of input signal segments into a more efficient parametric representation. The second method for bit rate reduction exploits the fact that for specific applications an exact reproduction of the input signal is not required and details of the signal which are regarded to be irrelevant for the receiver can be eliminated. Therefore this process is generally called irrelevancy reduction. Since in audio coding usually the highest possible subjective quality is desired, the irrelevancy reduction has to be based on perceptual characteristics of the human ear, e.g. by exploiting the so-called masking effect. Masking describes the increase of the audibility threshold for a soft test sound caused by a louder masking sound [1]. In a very simplified structure of an audio codec the encoder consists of an analysis block which is mainly based on a source model and of a quantization and coding block which is mainly based on a perception model (Figure 1a). The corresponding decoder first decodes the transmitted parameters and then synthesizes an audio signal which is presented to the receiver (Figure 1b).

Currently the two most commonly applied approaches are linear prediction based speech coding [2] on one hand and transform/subband coding for high quality audio [3] on the other hand. The first approach is mainly based on a physically motivated model of the signal source, i.e. the human vocal tract, while perceptual properties are usually modeled rather roughly. On the other hand the source models for high quality audio codecs are quite general and statistically motivated, while these codecs have sophisticated psychoacoustic models for

exploiting perceptual phenomena with respect to the audibility of quantization noise. In Section 2, a closer look is given to the underlying models of these two approaches which both produce at the decoder output an approximation of the original waveform. Section 3 gives an overview on coding techniques which are based on a parametric signal representation which no longer aims at the approximation of the original waveform. In order to obtain a very efficient representation of the audio signal, combinations of the different approaches seem to be useful. Some ideas for the extensions of source models and perception models allowing such combinations are given in Section 4, where also the approach taken for scaleable audio coding in MPEG-4 is described. A short summary and an outlook on possible approaches for future audio coding conclude this paper in Section 5.

2 Coding Techniques for Waveform Approximation

As well subband/transform audio codecs [3] as CELP (Code Excited Linear Prediction) type speech codecs [2] use perception models which describe the audibility of an error signal in relation to the original audio signal. The error signal which is caused by quantization is defined as the difference between the decoder output and the (delay compensated) encoder input signal. Therefore codecs of both types give an output signal which is an approximation of the input signal waveform. However the details of the source and perception models differ significantly as described in the following.

All commonly used low bit rate high quality audio codecs for general audio signals are based on a spectral decomposition by subband filtering or transform. Since no restriction of the input material is desired, no specific source model can be applied for such a codec. However the fact is exploited that normal audio signals use to be quasi stationary, i.e. show relatively constant signal properties within limited time intervals. Based on this assumption an analysis consisting of a segmentation into frames and subsequent mapping to the frequency domain results in a signal representation which is more efficient than the original PCM. Since the synthesis block in the decoder consists of a synthesis filter bank or inverse transform, errors introduced by quantization of a spectral component are limited to the corresponding frequency interval. This enables an efficient implementation of irrelevancy reduction by spectrally shaping the quantization noise by controlling the quantizer stepsizes for the individual spectral components. The necessary control information is obtained from a perception model (psychoacoustic model) exploiting frequency dependent masking effects in the human ear. Since this type of codec is based on time-to-frequency mapping, it is referred to as T/F codec in the following. The basic block diagrams of T/F encoder and decoder are shown in Figure 2. Investigations on the spectral resolution showed that a good choice is a frequency resolution around 20 Hz corresponding to a temporal resolution around 25 ms [4]. Based on refinements of source model and perception model several extensions to T/F coding have been introduced, e.g:

- the amplitude distribution of the spectral components can be exploited by Huffman coding [5] or vector quantization [6]
- the behavior in case of instationarities, especially attacks, can be enhanced by adaptively switching to higher temporal and lower spectral resolution [7] and by temporal noise shaping [8]
- longer term stationarities can be exploited by adaptive prediction for the spectral components [9]
- spectral envelopes can be transmitted efficiently by LPC based parameters [6].

T/F codecs show very good performance for coding music signals of all complexity levels at bit rates down to 16 kbit/s. At 16 kbit/s however speech codecs perform clearly better in case of speech signals due to their specialized source model.

Many of the recently designed speech codecs are based on CELP coding or similar techniques [2]. The fact that the type of input signals for which these codecs are designed is restricted to speech is reflected in a sophisticated, physically motivated source model. Speech generation is decomposed to an excitation which can be either voiced or unvoiced and to an acoustic tube which shapes the spectral envelope. This is reflected in the synthesis block of a speech decoder which usually consists of an excitation generator and a LPC synthesis filter (Figure 3b). In the corresponding encoder the filter parameters are derived from the spectral envelope of the input signal using LPC techniques and the optimum excitation is selected from a codebook in an analysis/synthesis loop (Figure 3a) by minimizing the residual error. The codebook often consists of two parts, an adaptive codebook and a stochastic codebook. The adaptive codebook contains previous samples which can be accessed with a time offset in order to model voiced signal parts, while the stochastic codebook mainly has the purpose of modeling unvoiced signal parts. In the loop a perception model can be taken into account by spectrally weighting the error between the synthesized signal and the input signal. However, compared to the sophisticated psychoacoustic models used in T/F codecs, this weighting filter does not exactly model the masking but is derived from the spectral envelope just in order to let the resulting error spectrum have a similar shape. It is interesting to note that the LPC synthesis filter length usually corresponds quite well to the propagation time of sound in the vocal tract. The described codec can be extended in correspondence to a refinement of the source model:

- vector quantization can be used for efficient coding of LPC parameters, e.g. exploiting the typical characteristics of the spectral envelope of speech
- other excitation models can be applied
- systems with reduced encoder complexity can be obtained with a pitch predictor outside of the analysis/synthesis loop.

CELP type speech codecs are used in different standards and show very good performance in the bit rate range between 4 and 24 kbit/s as long as the input material is speech. If they are applied to music, it becomes obvious that they are based on a specialized source model which is often not appropriate for general audio signals.

3 Coding Techniques without Waveform Approximation

For many audio signals the human ear is insensitive to changes of the waveform as long as certain spectral characteristics are maintained. Therefore coding techniques have been developed which are based on an efficient representation of the spectral characteristics without providing the capability to produce an approximation for the waveform of the input signal. The underlying source model for these techniques is based on the assumption that the input signal is composed of sinusoidal signal components and noise like components with relatively constant characteristics for certain time intervals [10]. The corresponding analysis has to extract for frames of input samples parameters that describe sinusoidal components (e.g. frequency, amplitude and phase) and noise components (e.g. spectral envelope). Due to the perception model phase information for the sinusoidal component needs not to be transmitted. Furthermore the transmission of the spectral envelope for the noise component is sufficient and no residual signal needs to be transmitted. If the input signal is restricted to speech, the source model can be refined in a way that there is either a voiced harmonic component which can be described by its pitch and spectral envelope or an unvoiced component again described by its spectral envelope [11]. A mixture of both types of signal components is also supported in several implementations [12]. If mixtures of speech and music are also allowed it is useful to extend this scheme in a way that harmonic components, individual sinusoidal components and noise components are supported [13]. A codec based on

this technique is in the following referred to as PARA codec, since it is rather based on a parametric signal representation than on waveform approximation. Figure 4a shows an encoder block diagram for a PARA encoder in which the coding of the model parameters is controlled by a perception model. The corresponding decoder consists of a parameter decoding unit and synthesizer units for the three types of signal components, i.e. sine wave generators for harmonic and individual sinusoidal components and noise generator with an appropriate filter for shaping the spectral envelope (Figure 4b). One problem which currently is only partly solved is the automatic separation of the input signal into the appropriate components.

Another difficulty in these approaches is the appropriate integration of a perception model. A psychoacoustic model as applied in T/F coding is not useful here, since first no difference signal can be calculated and second the criterion is not to subjectively minimize quantization noise. Therefore in the first implementations the perceptual criteria were only used to define the required accuracy for the different model parameters. However, recently an approach has been presented that orders individual sinusoidal components according to their relevance for perception by iteratively selecting components with amplitudes exceeding a masking threshold derived from components previously selected in the iteration [14]. This allows the selection of the most relevant components to be transmitted if their number is limited by bit rate constraints. This approach might be extended in order to select the parameters in a way that the excitation pattern in the human ear approximates the excitation patterns generated by the input signal as closely as possible. This would require more sophisticated models of perception than currently available, in order to correctly evaluate the influence of tones and noise.

PARA codecs have shown good performance for coding speech at bit rates down to 2 kbit/s and for music with low to medium complexity sound textures at bit rates between 6 and 16 kbit/s. Furthermore the applied synthesis methods allow independent modification of pitch and speed for playback by simply modifying the corresponding parameters during synthesis. For changing the pitch of speech signals it is interesting that the spectral envelopes can be maintained [11] in order to avoid effects such as the “shrinking head phenomenon“ which would be caused by shifting the formants towards higher frequencies in case of playback with increased pitch.

4 Combination of Different Coding Techniques

The different coding techniques described in Sections 2 and 3 are suited for different types of input material and different bit rates. Therefore it seems to be useful for internet and multimedia applications to support the different techniques within one unifying framework. This is the approach taken in the current ISO/IEC standardization activities within MPEG-4. The current status of MPEG-4 Audio is reflected by the so-called Verification Model (VM) described in [13] which is a software implementation of the tools and algorithms prepared for being included in the future standard to be issued in November 1998. The basic block diagram of the VM encoder (Figure 5) shows that the three techniques described in Sections 2 and 3 are supported. Although the three codec cores are already implemented there is no automatic separation or switching which would distribute the signal or components of it to the appropriate cores. However MPEG will not standardize the encoder but rather the syntax and the decoding procedure necessary to allow a flexible and efficient object oriented signal representation. Therefore a wide field for research and development of optimized encoders is still open.

In addition to coding efficiency other functionalities either providing more flexible access or allowing interactive manipulation seem to be very important for internet and multimedia applications. One of these functionalities is the so-called scaleability which allows decoding

of a subset of coded data. In several applications it is only necessary to have scaleability in a few large steps, e.g. for allowing either ISDN or modem/analog telephone line access. In this case the goal is to be more efficient than so-called simulcast, i.e. multiple transmission with different bit rates. In other applications, e.g. transmission with the possibility to assign priorities to packets, it would be a great advantage if scaleability in several small steps is possible. This would allow to provide a quality which gracefully degrades with reduced channel capacity or increased error rate. Since scaleability as well in large as in small steps is also regarded to be important for MPEG-4 Audio, the VM is meanwhile extended as shown in Figure 6. Small step scaleability is achieved by enhancement modules which encode the transmitted parameters in a way that with increasing bit rate the quantization artifacts are reduced and/or the bandwidth is increased. For large step scaleability there is the possibility to combine different cores. The most useful combination seems to be a CELP or PARA core operating at a low bit rate and the T/F core for enhancement, since as stated above the T/F core is well suited for high quality audio coding.

While in case of small step scaleability the coding efficiency for the highest possible bit rate usually does not reach that of a non-scaleable codec it is interesting to notice that large step scaleability might even increase the coding efficiency. For example some results within the MPEG-4 development have been reported which indicate that for speech signals a scaleable T/F-CELP combination performed better than a comparable pure T/F codec at the same total bit rate. Coming back to the considerations on source and perception models this result is no longer too surprising, since by combining the cores in an intelligent way [15] the advantages of the sophisticated source model of the CELP core can be combined with the advantages of the psychoacoustic model used in the T/F core. These results are an indication that in the future more efficient audio codecs might be developed by taking into account even more sophisticated source models. E.g. for music signals it might be worth considering techniques used or currently developed for music synthesis, e.g. physical modeling.

As stated above it is also desired to add features to audio coding allowing more interactivity. One example might be the construction of a new scene from several individually coded “sound objects“. Here the PARA core provides the possibility of manipulating playback speed and pitch, which might be used for the alignment or tuning of the different objects. Such applications would additionally be supported by the fact that PARA codecs are quite efficient for signals with low to medium complexity sound textures thus providing the capability of having many individually coded objects at reasonable total bit rates.

Even more interactivity and flexibility is possible if synthetic audio signals are taken into consideration. Therefore within MPEG an activity called “Synthetic Natural Hybrid Coding“ (SNHC) was started considering representation of synthetic video and audio. From SNHC there came two major contributions concerning audio.

The first contribution is a Text-To-Speech (TTS) system [13], [16] which specifies a syntax and a decoder for the combination of text, auxiliary information and control commands. This allows the integration of TTS synthesizers for different languages and supports interfaces to other modules like face animation, animated pictures and moving picture dubbing. TTS synthesizers however will not be specified by MPEG.

The second audio contribution from SNHC is called “Structured Audio“ [13], [17] and provides a specification of synthetic music/audio by a description separated into two parts: The first part is the so-called orchestra language which enables a downloadable definition of synthesis methods as instruments supporting various synthesis techniques. The second part is a score event syntax in which the sound generation can be controlled, e.g. invoke the instruments to play a tune. It is important to state here that this syntax also supports MIDI and General MIDI (GM) and provides interfaces for user interaction like keyboard, pedal or microphone. Compared to plain MIDI/GM the approach of SA has the advantage that the

content providers are now enabled to add to the scores (e.g. tunes) the specification of the synthesis. Thus they have control over the sound quality produced by the decoder. Since the tools used for synthesis also can be used for effects processing like reverb or 3D effects, these tools are also made available to the other cores. Therefore this core is called structured audio and effects processor (SAFX).

Figure 7 shows the block diagram of the MPEG-4 Audio VM decoder with the components for natural and synthetic audio, which in the full implementation provides real "SNHC".

5 Conclusions

This paper gives an overview on different approaches for exploiting source and perception models in audio coding. For the three codec types T/F, CELP and PARA the underlying models are explained. Due to the behavior for different types of input material and different bit rates a motivation for integrating different codec types into a unified framework like that of the MPEG-4 Audio Verification Model is given. New functionalities which will be provided besides coding efficiency for future interactive internet or multimedia applications are for example pitch change, speed change or scaleability. The way different codec types are combined for obtaining scaleability gives a motivation for future research in the field of low and very low bit rate audio coding. Automatic decomposition of a complex input signal into components for which individual coding is more efficient than direct coding of the mixture might be an interesting research area and is likely to have the potential for future improvements.

6 References

[1] E. Zwicker, H. Fastl: "Psychoacoustics, Facts and Models", Springer, Berlin - Heidelberg - New York, 1990.

[2] P. Noll: "Speech Coding for Communications", Proc. Eurospeech 1993.

[3] K. Brandenburg, 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, Jan./Feb. 1997.

[4] D.J. Meares, S.-W. Kim: "NBC Time/Frequency Module Subjective Tests: Overall Results", ISO/IEC JTC1/SC29/WG11 N0973, July 1995.

[5] K. Brandenburg: "OCF - A New Coding Algorithm for High Quality Sound Signals", Proc. IEEE ICASSP 1987, pp. 141-144.

[6] N. Iwakami, T. Moriya: "Transform-Domain Weighted Interleave Vector Quantization (TwinVQ)", 101st AES Convention, Preprint # 4377, Los Angeles 1996.

[7] B. Edler: "Coding of Audio Signals with Overlapping Transform and Adaptive Window Shape" (in German), Frequenz, Vol. 43, No. 9, pp. 252-256, Sept. 1989.

[8] J. Herre: "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)", 101st AES Convention, Preprint # 4384, Los Angeles 1996.

[9] H. Fuchs: "Improving MPEG Audio Coding by Backward Adaptive Linear Stereo Prediction", 99th AES Convention, Preprint # 4086, New York 1995.

[10] R.L. McAulay, T.F. Quatieri: "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE Trans. on ASSP, Vol. ASSP-34, No. 4, Aug. 1986.

[11] M. Nishiguchi, J. Matsumoto, S. Omori, K. Iijima: "Technical description of Sony IPC's proposal for MPEG-4 Audio and Speech Coding", ISO/IEC JTC1/SC29/WG11 MPEG95/0321, Nov. 1995.

[12] M. Nishiguchi, K. Iijima, J. Matsumoto: "Use of Mixed Voiced mode for 2.0Kbps parametric core of the VM", ISO/IEC JTC1/SC29/WG11 MPEG96/1408, Nov. 1996.

[13] MPEG Audio Subgroup: “Text of MPEG-4 Audio Working Draft V 3.0“, ISO/IEC JTC1/SC29/WG11 N1631, Apr. 1997.
[14] B. Edler, H. Purnhagen, C. Ferekidis: “ASAC - Analysis Synthesis Codec for Very Low Bit Rates“, 100th AES Convention, Preprint # 4179, Copenhagen 1996.
[15] B. Grill: “Implementation of a Scalable Audio Coder Based on AAC Modules and a Core Coder“, ISO/IEC JTC1/SC29/WG11 MPEG97/1816, Feb. 1997.
[16] M. Hahn, J. Lee: “MPEG-4 TTS Interface“, ISO/IEC JTC1/SC29/WG11 MPEG97/1524, Feb. 1997.
[17] E.D. Scheirer, B. Vercoe: “Synthetic Audio Specification Language Proposal for SNHC Audio“, ISO/IEC JTC1/SC29/WG11 MPEG97/1788, Feb. 1997.

Figures

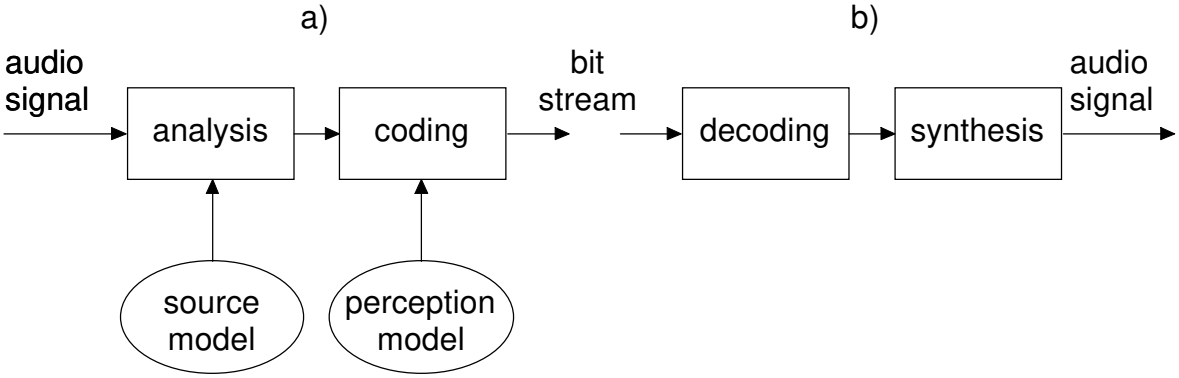

Figure 1: Simplified block diagram of audio encoder (a) and decoder (b)

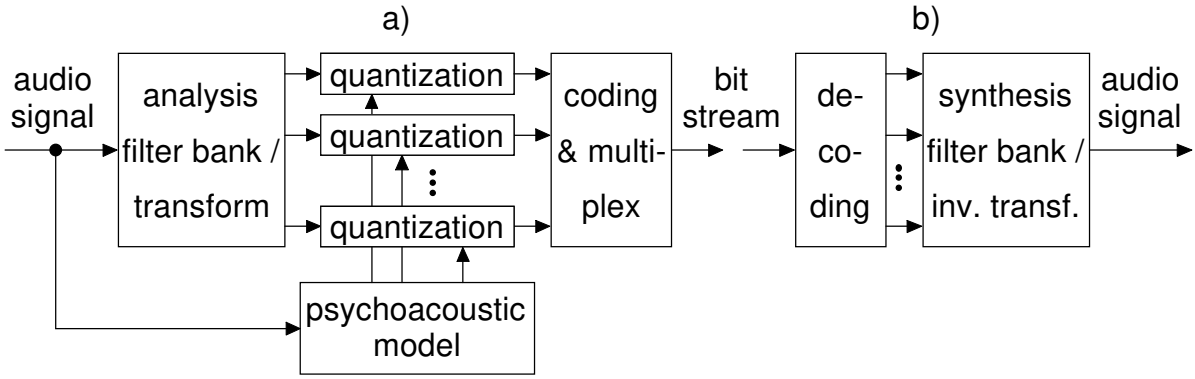

Figure 2: Basic block diagram of T/F encoder (a) and decoder (b)

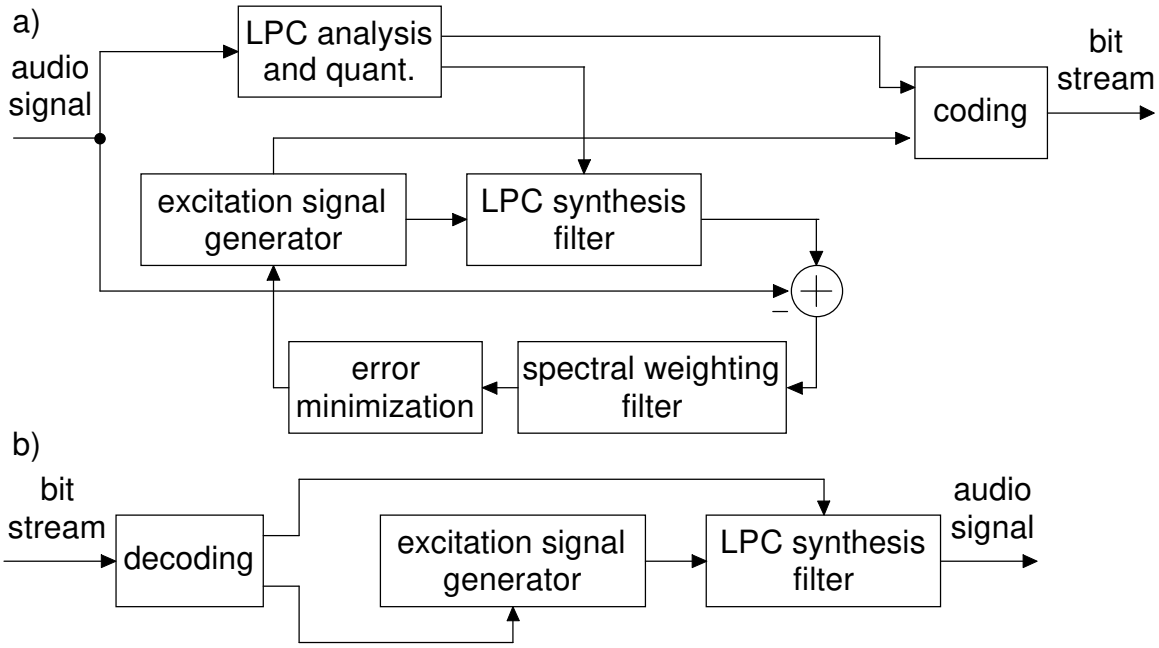

Figure 3: Block diagram of CELP encoder (a) and decoder (b)

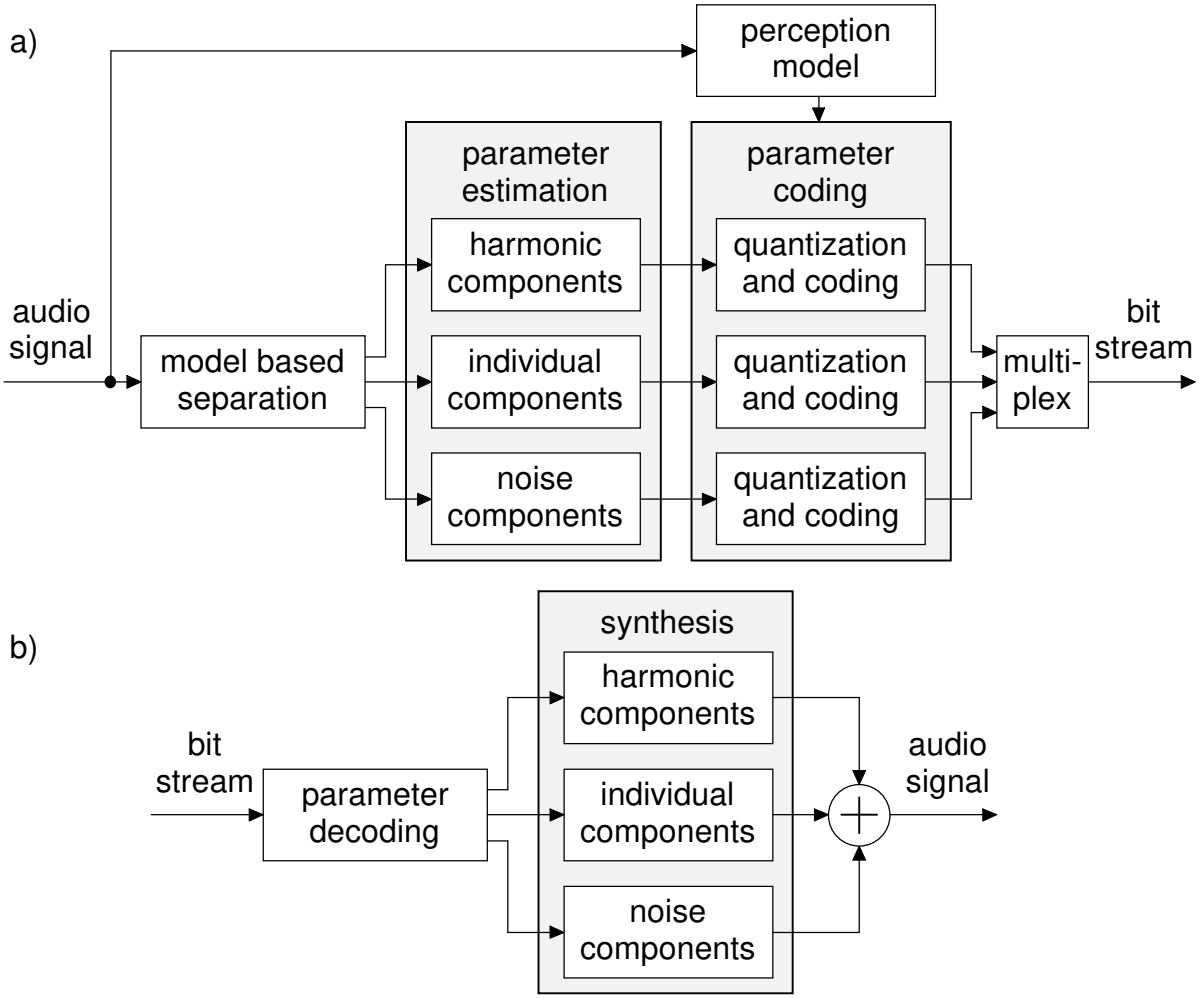

Figure 4: Block diagram of encoder (a) and decoder (b) for a codec based on a model with sinusoidal and noise components

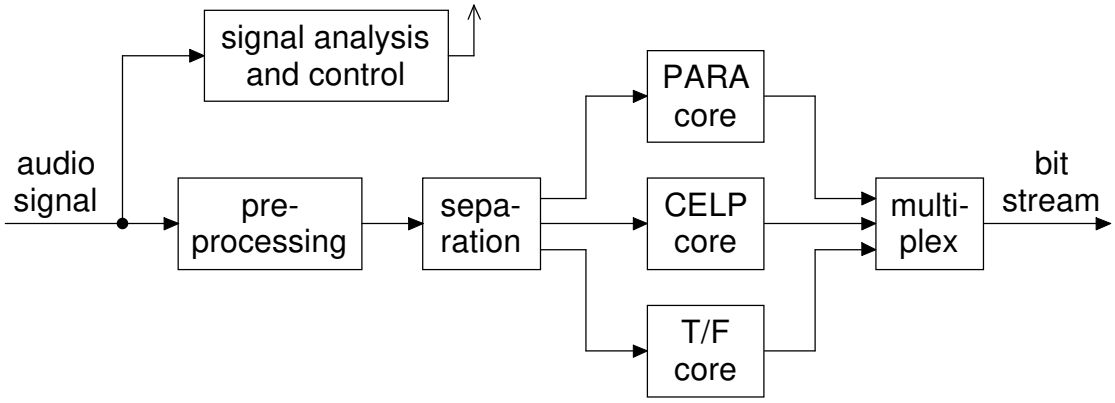

Figure 5: Basic encoder structure of the MPEG-4 Audio Verification Model

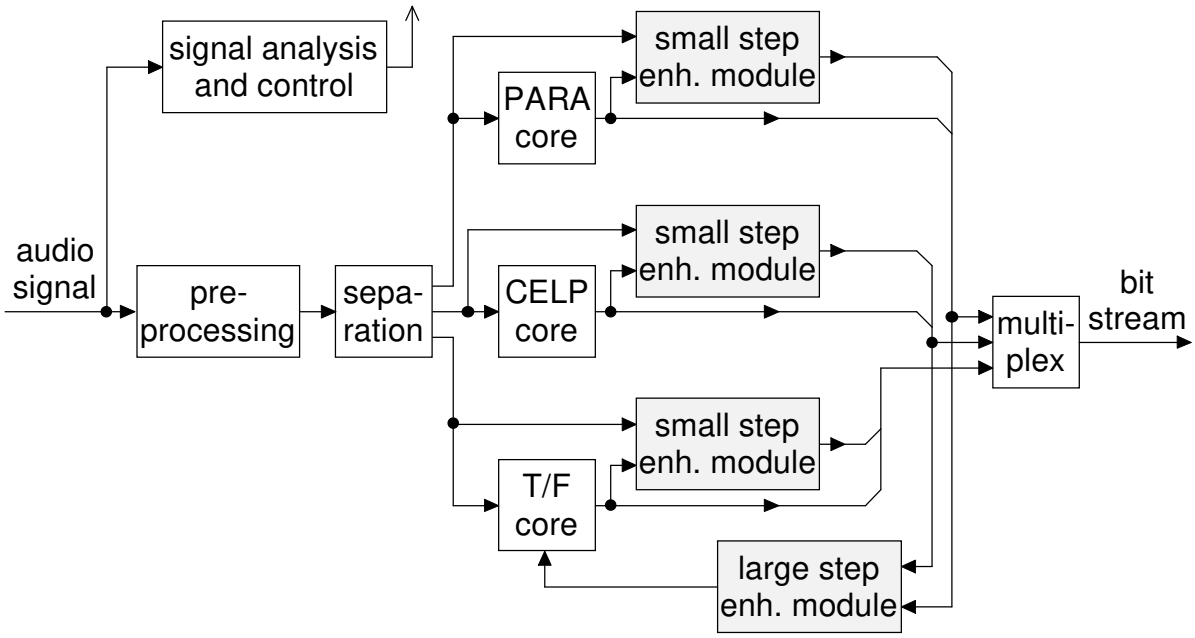

Figure 6: Block diagram of the MPEG-4 Audio VM encoder with modules for scaleability

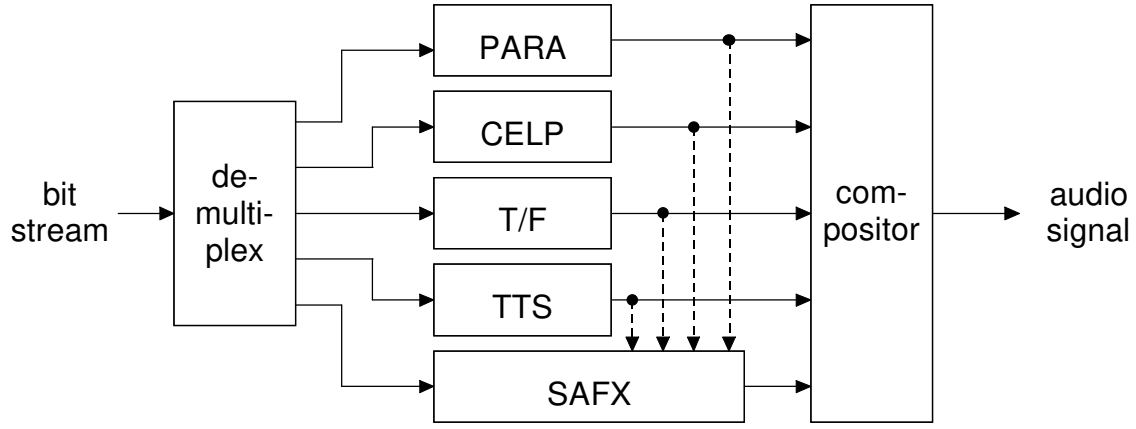

Figure 7: MPEG-4 Audio VM decoder for “Synthetic Natural Hybrid Coding“