Coding of Natural Audio in MPEG-4

Schuyler R. Quackenbush
AT&T Laboratories
180 Park Avenue
Florham Park, NJ, 09732, USA

ABSTRACT

MPEG-4 standardizes natural audio coding at bitrates ranging from 2 kbit/s, suitable for intelligible speech coding, to 64 kbit/s per channel, suitable for high-quality audio coding. Within this range, three categories of coding are defined: parametric coding, Code Excited Linear Predictive coding (CELP) and time/frequency (T/F) coding. The unique contribution of MPEG-4 audio is that not only does it scale across a wide range of bitrates, but it also scales across a broad set of other parameters, such as sampling rate, bandwidth, voice pitch and complexity. This paper presents an overview of the MPEG-4 natural audio coding framework and each of its component coding techniques.

1. INTRODUCTION

MPEG-4 audio coding integrates representations of natural audio, such as speech and audio coding systems, with representations of synthetic audio, such as MIDI and Text-to-Speech systems [1]. Natural audio representation systems are comprised of both analysis and synthesis components, (i.e. encoder and decoder) and attempt to reconstruct an existing audio signal under constraints, such as the bitrate of the intermediate representation. Synthetic audio representations are typically comprised of only the synthesis component. This paper will give an overview of natural audio coding in MPEG-4.

1.1 Core Tools

MPEG-4 standardizes natural audio coding at bitrates ranging from 2 kb/s per channel up to 64 kb/s per channel. To support such a broad range of rate and to achieve the highest audio quality within that range, three types of coders, or coding tools, have been defined. At the lowest rate, from 2 to 6 kbit/s, parametric coding is used, which is most effective for speech sampled at 8 kHz. At medium rates, from 6 to 24 kbit/s, Code Excited Linear Predictive (CELP) coding is used, which is able to support both speech and audio at sampling rates of 8 and 16 kHz. At the highest rates defined in MPEG-4, from 16 to 64 kbit/s, time/frequency coding techniques are used, such as the MPEG-2 Advanced Audio Coding standard (AAC). This coder can represent arbitrary audio signals with sampling rates from 8 to 96 kHz. The range of each coding tool as a function of bitrate and sampling frequency is shown in Figure 1.

1.2 Coding Tool Capabilities

A fundamental concept in MPEG-4 is that coding tools support more than just bitrate reduction.. Over the range of bitrates defined, the various tools support signals and signal qualities ranging from intelligible speech to high quality multichannel audio. Other functionalities supported by MPEG-4 coding tools are:

- Speed change, which allows the time scale to be altered without altering the pitch.
- Pitch change, which allows the pitch to be altered without altering the time scale.
- Bitrate scalability, which allows a bitstream to be parsed into a bitstream of lower rate which can still be decoded into a meaningful signal. The bitstream parsing can occur either during transmission or in the decoder.
- Bandwidth scalability, which is a particular case of bitrate scalability, in which part of a bitstream representing a part of the frequency spectrum can be discarded during transmission or decoding.
- Encoder complexity scalability, which allows encoders of different complexity to generate valid and meaningful bitstreams.
- Decoder complexity scalability, which allows a given bitstream to be decoded by decoders of different levels of complexity. The audio quality, in general, is related to the complexity of the encoder and decoder used.
- Error robustness, which provides the ability for a decoder to avoid or conceal audible distortion caused by transmission errors.

These functionalities are applicable to the individual coding tools (parametric, CELP and time/frequency) as well as across the coding tools.

To allow for smooth transitions in the reconstructed signal quality as the coding parameters are varied (i.e. bitrates bandwidth speed, etc.) a general framework for natural audio coding is defined, which is illustrated in Figure 2.

Starting with a coder operating at a low bitrate, by adding *enhancements*, both the coding quality and the audio bandwidth can be improved. These enhancements can be realized within a single coder or alternatively by combining different coding techniques.

The MPEG-4 system layer allows for profiles as a means to accommodate existing standards. Therefore MPEG-4 allows the use of several highly optimized coders, such as those standardized by the ITU-T, such that these coders can operate in a stand-alone mode with their own bitstream syntax.

2. OVERVIEW OF CODING TOOLS

2.1 Parametric Coding

The parametric coder core provides two sets of tools. The HVXC (Harmonic Vector eXcitation Coding) tool [2] allows coding of speech signals at 2 kbit/s and 4 kbit/s in a scalable way. HVXC also provides variable bit rate decoding at a typical average bit-rate of 1.5 kbps. The HILN (Harmonic and Individual Line plus Noise) tool [3] allows coding of non-speech signals such as music at bit rates of 4 kbit/s and higher. Both sets of tools allow for speed change and pitch change during decoding and can be combined to handle a wider range of signals and bit rates. Combining the output from the two coding tools to form an integrated parametric coder can be done by choosing only one of the two coders, dynamically switching between one of the two coders, or mixing the output from each of the two coders. To avoid hard transitions at frame boundaries when the HVXC or HILN decoders are switched on or off, the respective decoder output signals are faded in and out smoothly.

2.1.1 HVXC Decoder

The HVXC decoding process is composed of four steps: inverse quantization of parameters, generation of excitation signals for voiced frames by sinusoidal synthesis (harmonic synthesis) and noise component addition, generation of excitation signals for unvoiced frames by codebook look-up, and LPC synthesis. Spectral post-filtering is also used to enhance the synthesized speech quality.

The HVXC coder is capable of scalable delay, in which the encoder and decoder can independently select either low or normal delay modes. In both modes frame length is 20ms. In the encoder, algorithm delay can be selected to be either 26ms or 46ms. When 46ms delay is selected, one frame look ahead is used for pitch detection. When 26ms delay is selected, only the current frame is used for pitch detection. For both cases, syntax is common, all the quantizers are common, and bitstreams are compatible. In the decoder, algorithm delay can be selected to be either 10ms or 7.5ms. When 7.5ms delay is selected, the decoder frame interval is shifted by 2.5ms compared with the 10ms delay mode. In this case, excitation generation and LPC synthesis phase is shifted by 2.5 ms. Again, for both cases, syntax is common, all the quantizers are common, and bitstreams are compatible.

2.1.2 HILN Decoder

The HILN coding tool supports scalable speed, pitch, bitrate and complexity. The individual line basic decoder reconstructs the line parameters *frequency*, *amplitude*, and *envelope* from the bitstream. The enhanced decoder reconstructs the line parameters *frequency*, *amplitude*, and *envelope* with finer quantization and additionally reconstructs the line parameters *phase*.

Due to the phase continuation in the individual line synthesis process, the speed of the decoded signal can be changed by simply changing the frame length without any other modifications. The ratio of the encoder frame length to the decoder frame length directly corresponds to the speed-up factor. Additionally, the pitch of the decoded signal can be varied without affecting the frame length and without causing phase discontinuities. Pitch change is performed by multiplying each frequency parameter by a factor before it is used in the synthesis process.

The enhanced decoder augments the basic decoder by both reconstructing the line parameters with finer quantization and also by decoding the line phases. The latter allows the generation of a signal that approximates the coder input waveform. The additional information required for this approximation is contained a separate enhancement bitstream, thus permitting bitrate and complexity scalability in the HILN decoder.

If noise parameters are transmitted in a given frame, a noise signal with the indicated spectral shape is synthesized and added to the audio signal generated by the harmonic and individual line synthesizers.

2.2 CELP Coding

The CELP decoder consists of an excitation source and a synthesis filter and, optionally, a postfilter [4]. The excitation source has both periodic components, contributed by an adaptive codebook, and random components contributed by one or more fixed codebooks. At the decoder, the excitation signal is reconstructed using the codebook indices (pitch lag for the adaptive codebook and shape index for the fixed codebook) and gain indices (adaptive and fixed codebook gains). This excitation signal is then filtered by the linear predictive synthesis filter (LP synthesis filter). This filter is obtained by interpolating the LPC coefficients of successive analysis frames, where the LPC coefficients are reconstructed using the LPC indices. Finally, a postfilter is applied in order to enhance the speech sound quality. Two sampling rates are supported: 16 and 8 kHz.

Bitrate scalability is possible for the 8 kHz sampling rate by adding enhancement layers, which can be added in steps of 2kb/s. A maximum of 3 enhancement layers may be added, giving the possibility to add 2, 4 or 6 kbit/s to the base bit rate. Variable bitrates are also supported.

When using 16 kHz sampling rate, it is possible to decode speech by using only a part of the bitstream, thereby giving complexity scalability. Other methods of complexity scaling that are supported are: simplified LPC interpolation, postfilter present or absent, and reduced order LPC synthesis. These instances of complexity scalability are strictly implementation dependent and do not depend on the bitstream syntax. In the case of a software decoder, the complexity level can even be changed during run-time so as to cope with the computational load in a limited-capacity terminal or in a multi-tasking environment.

Bandwidth scalability to cover both sampling rates is realized by adding a bandwidth extension tool to the CELP coder. This is an additional tool, supported in the 8 kHz sampling rate mode only, which may be added if scalability to 16 kHz

sampling rate is required. This tool is distinct from the 16 kHz sampling rate mode.

2.3 T/F Coding

A simplified block diagram of the MPEG-4 time/frequency decoder is show in Figure 3. At 64kb/s per channel, it is exactly MPEG-2 AAC [5], which at this signal compression has demonstrated excellent audio quality. At lower rates, additional tools can be used to gain features such as bitrate or bandwidth scalability or error robustness.

2.3.1 Advanced Audio Coding

MPEG-2 AAC is the core of MPEG-4 time/frequency coding. It uses a high frequency-resolution, 1024-band filterbank for maximum statistical signal gain, but can increase its time resolution by switching to 128 bands when the signal exhibits non-stationarity. This resolution-switching, or "block switching" capability serves to contains the backward spread of quantization noise in the time domain. In an even more flexible way the temporal noise shaping (TNS) tool controls the time-domain aspects of quantization noise. Backward adaptive prediction is (optionally) used on each spectral coefficient as a means to effectively increase the resolution of the filterbank.

Spectral coefficients are quantized using one non-uniform quantizer per *scale factor band*, a division of the set of coefficients that approximates the critical band structure. The encoder psychoacoustic model sets the quantizer stepsize such that quantization noise is masked by the signal. In the noiseless coding tool, the sets of quantized coefficients associated with each scale factor band are grouped into *sections*, with an integral number of scale factor bands per section. The quantized coefficients are Huffman coded in 2- or 4-tuples using one codebook per section.

With multichannel signals AAC can use M/S (sum/difference) coding or intensity stereo coding, in which M/S or intensity can adapt at each scale factor band for every transform block. It also defines a coupling channel tool, which is a generalization of intensity stereo both in terms of the number of channels jointly coded and the technique of representing the joint signal.

The coder can operate in a constant-quality/variable-rate mode or in a constant-rate mode. In the latter case, the quantizer stepsizes are adjusted in an interative manner until the target bitcount for the current block is achieved.

2.3.2 Other T/F tools

The bit-sliced arithmetic coding (BASC) tool [6] is an alternative to the AAC noiseless coding tool, and provides a fine-grained scalability in bitrate, from 16 kbit/s to 64 kbit /s in 1 kbit /s steps. BASC encodes like-significance bits in the quantized coefficients per *coding band*, which is a set of 32 contiguous coefficients, and repeats this coding from most significant to least significant bit in each coding band. An advantage of this tool is that a bitstream in AAC format can be easily trans-codeded into one in BASC format.

The transform-domain weighted interleaved vector quantization (TwinVQ) tool [7] is an alternative to the AAC noiseless coding and quantization tools. It uses an LPC model to specify the quantizer stepsizes and vector quantization for the set of interleaved and quantized spectral coefficients. This tool is particularly suitable for systems that requires bitrate scalability or error robustness.

3. CONCLUSIONS

Not only does MPEG-4 natural audio coding provide signal compression across a wide range of bitrates, but it uniquely provides significant scalability for a number of key system parameters, such as channel bitrate, signal bandwidth, reconstructed signal time scale, voice pitch and decoder complexity. They key to this scalability is a coding system composed of a set of core coders, each suited to a segment of the bitrate range, and a scheme to combine their outputs, as appropriate to achieve parameter scaling both within a core coder and across the set of coders.

4. ACKNOWLEDGMENTS

I would like to thank the other authors of the MPEG-4 Committee Draft document (ISO/IEC CD 14496-3) for providing much of the material in this paper.

5. REFERENCES

1. R. Koenen, "MPEG-4 Overview," http://drogo.cselt.it/mpeg/public/w1909.htm.
2. M. Nishiguchi and J. Matsumoto, "Harmonic and Noise Coding of LPC Residuals with Classified Vector Quantization," Proc. IEEE ICASSP, May, 1995.
3. B. Edler, H. Purnhagen, C. Ferekidis, "ASAC-Analysis/Synthesis Audio Codec for Very Low-Bit Rates," 100th Conv. of AES, May, 1996, preprint 4179.
4. W. B. Kleijn, K. K. Paliwal, *Speech Coding and Synthesis*. Elsevier Science, Amsterdam, 1995
5. M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Diets, J. Herre, G. Davidson and Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," 101st Conv. of AES, Nov 1996, preprint 4382.
6. S. Park, Y. Kim, Y, Seo, "Multi-Layer Bit-Sliced Bit-Rate Scalable Audio Coding," 103rd Conv. of AES, Sep, 1997, preprint 4520.
7. T. Moriya, "Transform-Domain Weighted Interleave Vector Quantization (TwinVQ)," 101st Conv. of AES, Nov, 1996, preprint 4377.

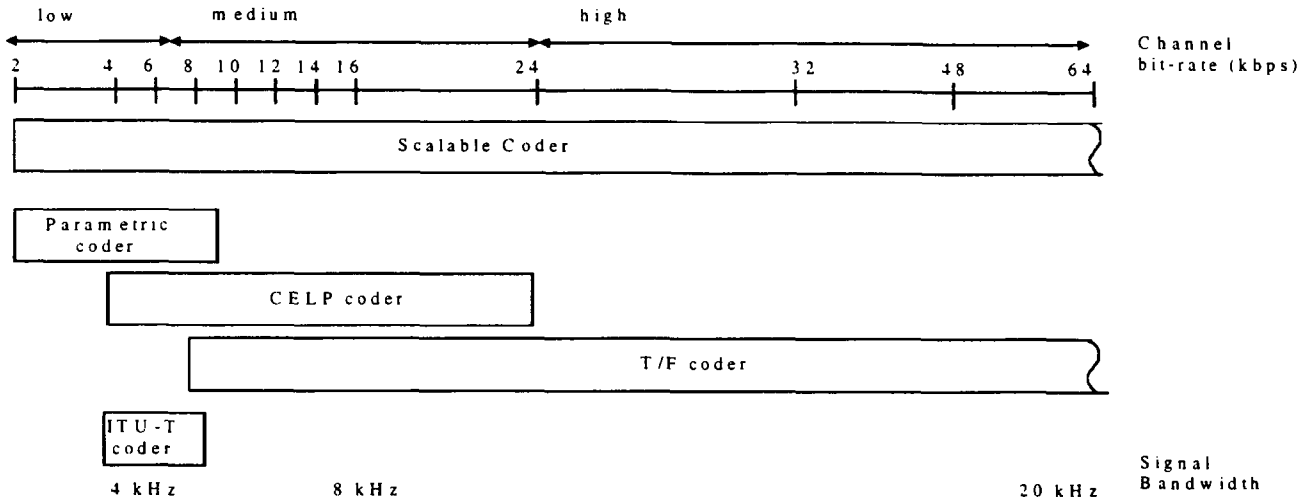

Figure 1. Range of bitrates and signal bandwidth for each of the MPEG-4 audio coding tools.

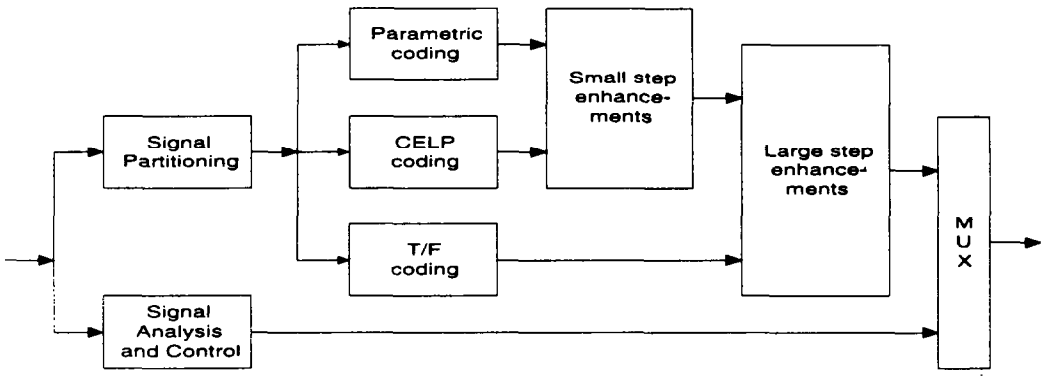

Figure 2. MPEG-4 natural encoder system block diagram.

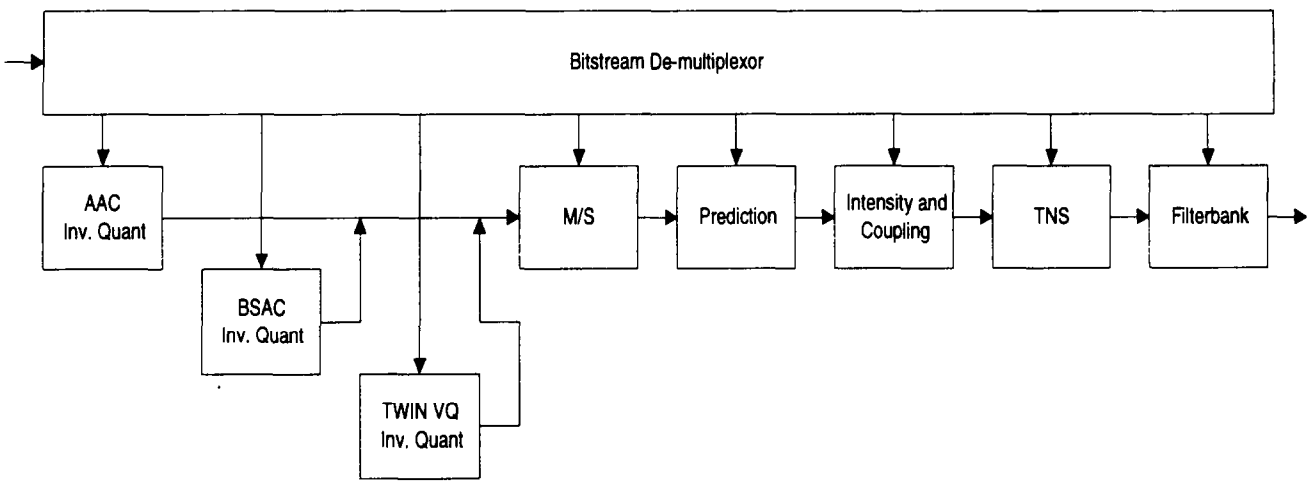

Figure 3. Block diagram of time/frequency decoder