ELSEVIER

Signal Processing: *Image Communication* 15 (2000) 423–444

SIGNAL PROCESSING: IMAGE COMMUNICATION

www.elsevier.nl/locate/image

MPEG-4 natural audio coding

Karlheinz Brandenburg[a,*], Oliver Kunz[a], Akihiko Sugiyama[b]

[a]*Fraunhofer Institut für Integrierte Schaltungen IIS, D-91058 Erlangen, Germany*
[b]*NEC C&C Media Research Laboratories, 1-1, Miyazaki 4-chome, Miyamae-ku, Kawasaki 216-8555, Japan*

Abstract

MPEG-4 audio represents a new kind of audio coding standard. Unlike its predecessors, MPEG-1 and MPEG-2 high-quality audio coding, and unlike the speech coding standards which have been completed by the ITU-T, it describes not a single or small set of highly efficient compression schemes but a complete toolbox to do everything from low bit-rate speech coding to high-quality audio coding or music synthesis. The natural coding part within MPEG-4 audio describes traditional type speech and high-quality audio coding algorithms and their combination to enable new functionalities like scalability (hierarchical coding) across the boundaries of coding algorithms. This paper gives an overview of the basic algorithms and how they can be combined.

1. Introduction

Traditional high-quality audio coding schemes like MPEG-1 Layer-3 (aka .mp3) have found their way into many applications including widespread acceptance on the Internet. MPEG-4 audio is scheduled to be the successor of these, building and expanding on the acceptance of earlier audio coding formats. To do this, MPEG-4 natural audio coding has been designed to fit well into the philosophy of MPEG-4. It enables new functionalities and implements a paradigm shift from the linear storage or streaming architecture of MPEG-1 and MPEG-2 into objects and presentation rendering. While most of these new functionalities live within the tools of MPEG-4 structured audio and audio BIFS, the syntax of the "classical" audio coding algorithms within MPEG-4 natural audio has been defined and amended to implement scalability and the notion of audio objects. This way MPEG-4 natural audio goes well beyond classic speech and audio coding algorithms into a new world which we will see unfold in the coming years.

2. Overview

The tools defined by MPEG-4 natural audio coding can be combined to different audio coding algorithms. Since no single coding paradigm was found to span the complete range from very low bit-rate coding of speech signals up to high-quality multi-channel audio coding, a set of different algorithms has been defined to establish optimum coding efficiency for the broad range of anticipated applications (see Fig. 1 and [9]). The following list introduces the main algorithms and the reason for

* Corresponding author.
E-mail address: bdg@iis.fhg.de (K. Brandenburg)

0923-5965/00/$ - see front matter
PII: S0923-5965(99)00056-9

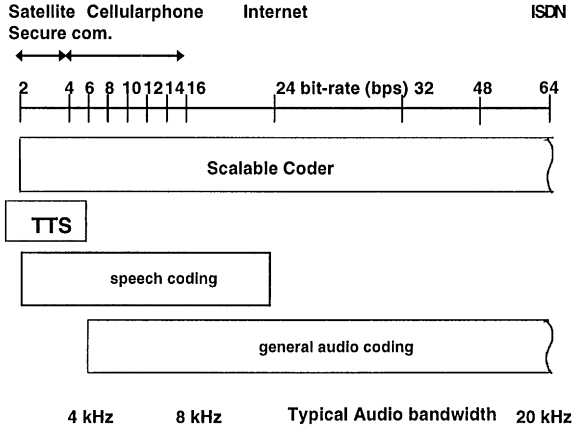


Fig. 1. Assignment of codecs to bit-rate ranges.

their inclusion into MPEG-4. The following sections will give more detailed descriptions for each of the tools used to implement the coding algorithms. The following lists the major algorithms of MPEG-4 natural audio. Each algorithm was defined from separate coding tools with the goals of maximizing the overlap of tools between different algorithms and maximizing the flexibility in which tools can be used to generate different flavors of the basic coding algorithms.

- HVXC Low-rate clean speech coder
- CELP Telephone speech/wideband speech coder
- GA General Audio coding for medium and high qualities
- TwinVQ Additional coding tools to increase the coding efficiency at very low bit-rates.

In addition to the coding tools used for the basic coding functionality, MPEG-4 provides techniques for additional features like bit stream scalability. Tools for these features will be explained in Section 7.

3. General Audio Coding (AAC-based)

This key component of MPEG-4 Audio covers the bit-rate range of 16 kbit/s per channel up to bit-rates higher than 64 kbit/s per channel. Using MPEG-4 General Audio quality levels between "better than AM" up to "transparent audio quality" can be achieved. MPEG-4 General Audio supports four so-called Audio Object Types (see the paper on MPEG-4 Profiling in this issue), where AAC Main, AAC LC, AAC SSR are derived from MPEG-2 AAC (see [2]), adding some functionalities to further improve the bit-rate efficiency. The fourth Audio Object Type, AAC LTP is unique to MPEG-4 but defined in a backwards compatible way.

Since MPEG-4 Audio is defined in a way that it remains backwards compatible to MPEG-2 AAC, it supports all tools defined in MPEG-2 AAC including the tools exclusively used in Main Profile and scalable sampling rate (SSR) Profile, namely Frequency domain prediction and SSR filterbank plus gain control.

Additionally, MPEG-4 Audio defines ways for bit-rate scalability. The supported methods for bit-rate scalability are described in Section 7.

Fig. 2 shows the arrangement of the building blocks of an MPEG-4 GA encoder in the processing chain. These building blocks will be described in the following subsections. The same building blocks are present in a decoder implementation, performing the inverse processing steps. For the sake of simplicity we omit references to decoding in the following subsections unless explicitly necessary for understanding the underlying processing mechanism.

3.1. Filterbank and block switching

One of the main components in each transform coder is the conversion of the incoming audio signal from the time domain into the frequency domain.

MPEG-2 AAC supports two different approaches to this. The standard transform is a straightforward modified discrete cosine transform (MDCT). However, in the AAC SSR Audio Object Type a different conversion using a hybrid filter bank is applied.

3.1.1. Standard filterbank

The filterbank in MPEG-4 GA is derived from MPEG-2 AAC, i.e. it is an MDCT supporting block

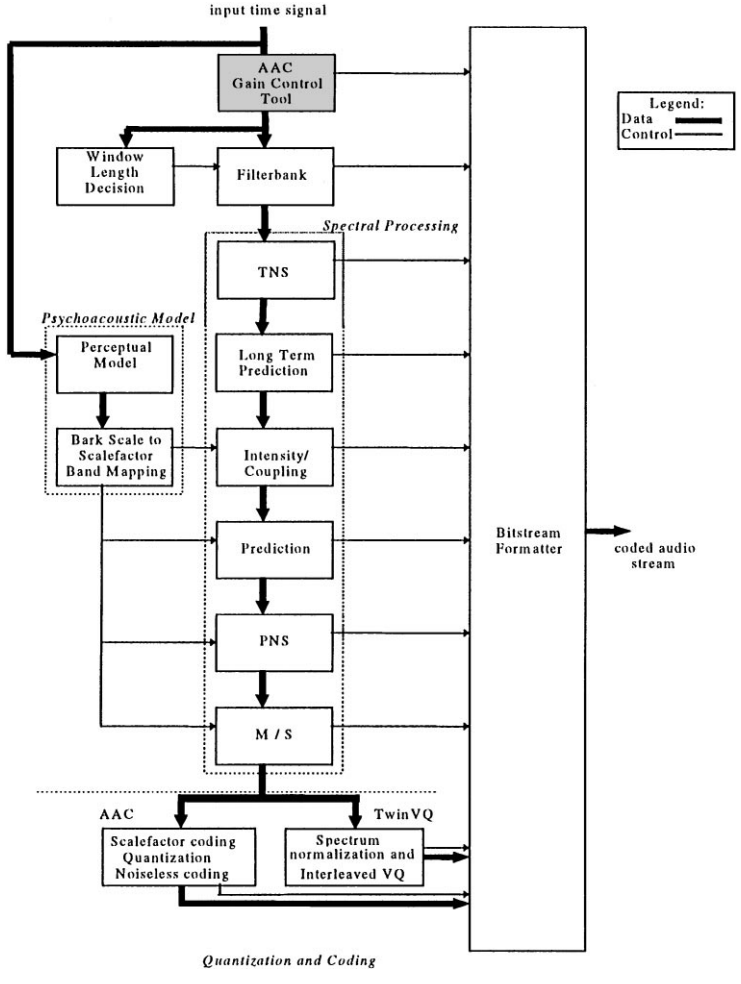


Fig. 2. Building blocks of the MPEG-4 General Audio Coder.

lengths of 2048 points and 256 points which can be switched dynamically. Compared to previously known transform coding schemes the length of the long block transform is rather high, offering improved coding efficiency for stationary signals. The shorter of the two block length is rather small, providing optimized coding capabilities for transient signals. MPEG-4 GA supports an additional mode with block lengths of 1920/240 points to facilitate scalability with the speech coding algorithms in MPEG-4 Audio (see VI-C). All blocks are overlapped by 50% with the preceding and the following block.

For improved frequency selectivity the incoming audio samples are windowed before doing the transform. MPEG-4 AAC supports two different window shapes that can be switched dynamically. The two different window shapes are a sine-shaped window and a Kaiser–Bessel derived (KBD) Window offering improved far-off rejection compared to the sine-shaped window.

An important feature of the time-to-frequency transform is the signal adaptive selection of the transform length. This is controlled by analyzing the short time variance of the incoming time signal.

To assure block synchronicity between two audio channels with different block length sequences eight short transforms are performed in a row using 50% overlap each and specially designed transition windows at the beginning and the end of a short sequence. This keeps the spacing between consecutive blocks at a constant level of 2048 input samples.

For further processing of the spectral data in the quantization and coding part the spectrum is arranged in the so-called scalefactor bands roughly reflecting the bark scale of the human auditory system.

3.1.2. Filterbank and gain control in SSR profile

In the SSR profile the MDCT is preceded by a processing block containing a uniformly spaced 4-band polyphase quadrature filter (PQF) and a Gain control module. The Gain control can attenuate or amplify the output of each PQF band to reduce pre-echo effects.

After the gain control is performed, an MDCT is calculated on each PQF band, having a quarter of the length of the original MDCT.

3.2. Frequency-domain prediction

The frequency-domain prediction improves redundancy reduction of stationary signal segments. It is only supported in the Audio Object Type AAC Main. Since stationary signals can nearly always be found in long transform blocks, it is not supported in short blocks. The actual implementation of the predictor is a second-order backwards adaptive lattice structure, independently calculated for every frequency line. The use of the predicted values instead of the original ones can be controlled on a scalefactor band basis and is decided based on the achieved prediction gain in that band.

To improve stability of the predictors, a cyclic reset mechanism is applied which is synchronized between encoder and decoder via a dedicated bitstream element.

The required processing power of the frequency-domain prediction and the sensitivity to numerical imperfections make this tool hard to use on fixed point platforms. Additionally, the backwards adaptive structure of the predictor makes such bitstreams quite sensitive to transmission errors.

3.3. Long-term prediction (LTP)

Long-term prediction (LTP) is an efficient tool for reducing the redundancy of a signal between successive coding frames newly introduced in MPEG-4. This tool is especially effective for the parts of a signal which have clear pitch property. The implementation complexity of LTP is significantly lower than the complexity of the MPEG-2 AAC frequency-domain prediction. Because the Long-Term Predictor is a forward adaptive predictor (prediction coefficients are sent as side information), it is inherently less sensitive to round-off numerical errors in the decoder or bit errors in the transmitted spectral coefficients.

3.4. Quantization

The adaptive quantization of the spectral values is the main source of the bit-rate reduction in all transform coders. It assigns a bit allocation to the spectral values according to the accuracy demands determined by the perceptual model, realizing the irrelevancy reduction. The key components of the quantization process are the actually used quantization function and the noise shaping that is achieved via the scalefactors (see III-E). The quantizer used in MPEG-4 GA has been designed similar to the one used in MPEG 1/2 Layer-3. It is a non-linear quantizer with an $x^{0.75}$ characteristic. The main advantage of this non-linear quantization over a conventional linear quantizer is the implicit noise shaping that this quantization creates. The absolute quantizer stepsize is determined via a specific bitstream element. It can be adjusted in 1.5 dB steps.

3.5. Scalefactors

While there is already an inherent noise shaping in the non-linear quantizer it is usually not sufficient to achieve acceptable audio quality. To improve the subjective quality of the coded signal the noise is further shaped via scalefactors. The way the scalefactors are working is the following: Scalefactors are used to amplify the signal in certain spectral regions (the scalefactor bands) to increase the signal-to-noise ratio in these bands. Thus they

implicitly modify the bit-allocation over frequency since higher spectral values usually need more bits to be coded afterwards. Like the global quantizer the stepsize of the scalefactors is 1.5 dB.

To properly reconstruct the original spectral values in the decoder the scalefactors have to be transmitted within the bitstream. MPEG-4 GA uses an advanced technique to code the scalefactors as efficiently as possible. First, it exploits the fact that scalefactors usually do not change too much from one scalefactor band to another. Thus a differential encoding already provides some advantage. Second, it uses a Huffman code to further reduce the redundancy within the scalefactor data.

3.6. Noiseless coding

The noiseless coding kernel within an MPEG-4 GA encoder tries to optimize the redundancy reduction within the spectral data coding. The spectral data is encoded using a Huffman code which is selected from a set of available code books according to the maximum quantized value. The set of available codebooks includes one to signal that all spectral coefficients in the respective scalefactor band are "0", implying that there are neither spectral coefficients nor a scalefactor transmitted for that band. The selected table has to be transmitted inside the so-called section_data, creating a certain amount of side-information overhead. To find the optimum tradeoff between selecting the optimum table for each scalefactor band and minimizing the number of section_data elements to be transmitted an efficient grouping algorithm is applied to the spectral data.

3.7. Joint stereo coding

Joint stereo coding methods try to increase the coding efficiency when encoding stereo signals by exploiting commonalties between the left and right signal. MPEG-4 GA contains 2 different joint stereo coding algorithms, namely mid-side (MS) stereo coding and Intensity stereo coding.

MS stereo applies a matrix to the left and right channel signals, computing sum and difference of the two original signals. Whenever a signal is concentrated in the middle of the stereo image, MS stereo can achieve a significant saving in bit-rate. Even more important is the fact that by applying the inverse matrix in the decoder the quantization noise becomes correlated and falls in the middle of the stereo image where it is masked by the signal.

Intensity stereo coding is a method that achieves a saving in bit-rate by replacing the left and the right signal by a single representing signal plus directional information. This replacement is psychoacoustically justified in the higher frequency range since the human auditory system is insensitive to the signal phase at frequencies above approximately 2 kHz.

Intensity stereo is by definition a lossy coding method thus it is primarily useful at low bit-rates. For coding at higher bit-rates only MS stereo is used.

3.8. Temporal noise shaping

Conventional transform coding schemes often encounter problems with signals that vary heavily over time, especially speech signals. The main reason for this is that the distribution of quantization noise can be controlled over frequency but is constant over a complete transform block. If the signal characteristic changes drastically within such a block without leading to a switch to shorter transform lengths, e.g. in the case of pitchy speech signals this equal distribution of quantization noise can lead to audible artifacts.

To overcome this limitation, a new feature called temporal noise shaping (TNS) (see [5]) was introduced into MPEG-2 AAC. The basic idea of TNS relies on the duality of time and frequency domain. TNS uses a prediction approach in the frequency domain to shape the quantization noise over time. It applies a filter to the original spectrum and quantizes this filtered signal. Additionally, quantized filter coefficients are transmitted in the bitstream. These are used in the decoder to undo the filtering performed in the encoder, leading to a temporally shaped distribution of quantization noise in the decoded audio signal.

TNS can be viewed as a postprocessing step of the transform, creating a continuous signal adaptive filter bank instead of the conventional two-step switched filter bank approach. The actual

implementation of the TNS approach within MPEG-2 AAC and MPEG-4 GA allows for up to three distinct filters applied to different spectral regions of the input signal, further improving the flexibility of this novel approach.

3.9. Perceptual noise substitution (PNS)

A feature newly introduced into MPEG-4 GA, i.e. not available within MPEG-2 AAC, is the perceptual noise substitution (PNS) (see [6]). It is a feature aiming at a further optimization of the bit-rate efficiency of AAC at lower bit-rates.

The technique of PNS is based on the observation that "one noise sounds like the other". This means that the actual fine structure of a noise signal is of minor importance for the subjective perception of such a signal. Consequently, instead of transmitting the actual spectral components of a noisy signal, the bit-stream would just signal that this frequency region is a noise-like one and give some additional information on the total power in that band. PNS can be switched on a scalefactor band basis so even if there just are some spectral regions with a noisy structure PNS can be used to save bits. In the decoder, a randomly generated noise will be inserted into the appropriate spectral region according to the power level signaled within the bit-stream.

From the above description it is obvious that the most challenging task in the context of PNS is not to enter the appropriate information into the bit-stream but reliably determining which spectral regions may be treated as noise-like and thus may be coded using PNS without creating severe coding artifacts. A lot of work has been done on this task, most of which is reflected in [20].

4. TwinVQ

To increase coding efficiency for coding of musical signals at very low bit-rates, TwinVQ-based coding tools are part of the General Audio coding system in MPEG-4 audio. The basic idea is to replace the conventional encoding of scalefactors and spectral data used in MPEG-4 AAC by an interleaved vector quantization applied to a normalized spectrum (see [10,11]). The rest of the processing chain remains identical as can be seen in Fig. 2.

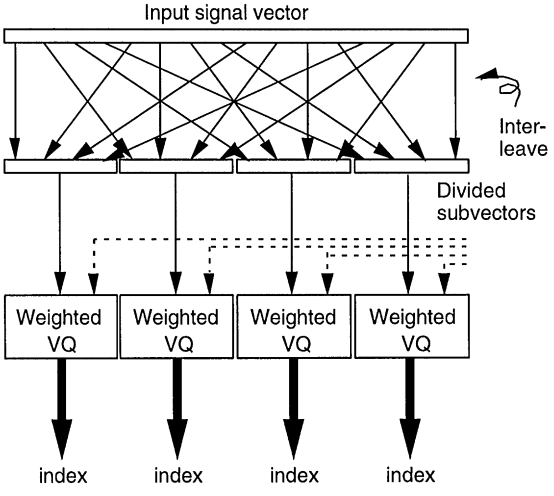


Fig. 3. Weighted interleave vector quantization.

Fig. 3 visualizes the basic idea of the weighted interleaved vector quantization (TwinVQ) scheme. The input signal vector (spectral coefficients) is interleaved into subvectors. These subvectors are quantized using vector quantizers.

Twin VQ can achieve a higher coding efficiency at the cost of always creating a minimum amount of loss in terms of audio quality. Thus, the break even point between Twin VQ and MPEG-4 AAC is at fairly low bit-rates (below 16 kbit/s per channel).

5. Speech coding in MPEG-4 Audio

5.1. Basics of speech coding [4,12]

Most of the recent speech coding algorithms can be categorized as a spectrum coding or a hybrid coding. Spectrum coding models the input speech signal based on a vocal tract model which consists of a signal source and a filter as shown in Fig. 4. A set of parameters obtained by analyzing the input signal are transmitted to the receiver.

Hybrid coding synthesizes an approximated speech signal based on a vocal tract model. A set of parameters used for this first synthesis are modified

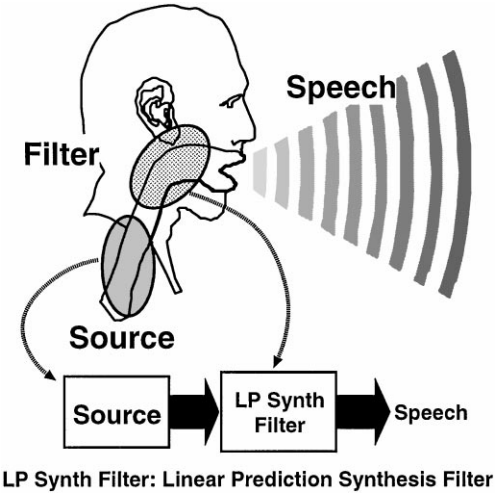


Fig. 4. Vocal tract model.

to minimize the error between the original and the synthesized speech signals. A best parameter set can be searched for by repeating this analysis-by-synthesis procedure. The obtained set of parameters are transmitted to the receiver as the compressed data after quantization. In the decoder, a set of parameters for Source and linear prediction (LP) synthesis filtering are recovered by inverse quantization. These parameter values are used to operate the same vocal tract model as in the encoder.

Fig. 5 depicts a block diagram of hybrid speech coding. Source and LP Synth Filter in Fig. 5 correspond to those in Fig. 4. Upon parameter search, the error between the input signal and the synthesized signal is weighted by a PW (perceptually weighted) filter. This filter has a frequency response which takes the human auditory system into consideration, thus a perceptually best parameter selection can be achieved.

5.2. Overview of the MPEG-4 Natural Speech Coding Tools

MPEG-4 Natural Speech Coding Tool Set [8] provides a generic coding framework for a wide range of applications with speech signals. Its bit-rate coverage spans from as low as 2–23.4 kbit/s. Two different bandwidths of the input speech signal

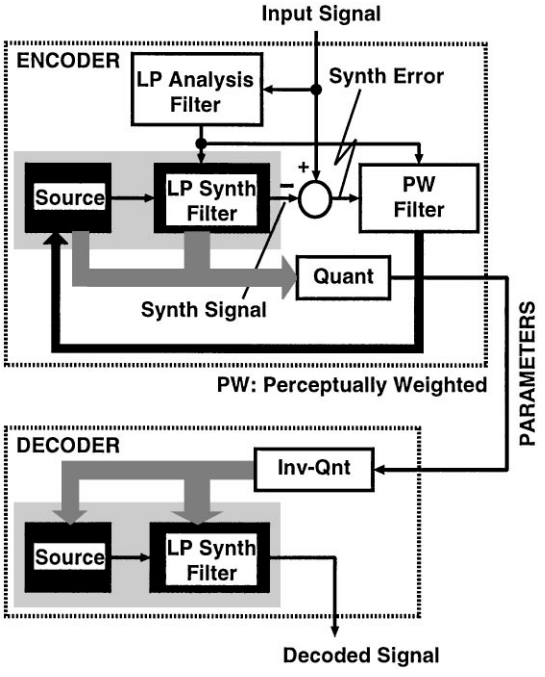


Fig. 5. Hybrid speech coding.

are covered, namely, 4 and 7 kHz. MPEG-4 Natural Speech Coding Tool Set contains two algorithms: harmonic vector excitation coding (HVXC) and code excited linear predictive coding (CELP). HVXC is used at a low bit-rate of 2 or 4 kbit/s. Higher bit-rates than 4 kbit/s in addition to 3.85 kbit/s are covered by CELP. The algorithmic delay by either of these algorithms is comparable to that of other standards for two-way communications, therefore, MPEG-4 Natural Speech Coding Tool Set is also applicable to such applications. Storage of speech data and broadcast are also promising applications of MPEG-4 Natural Speech Coding Tool Set. The specifications of MPEG-4 Natural Speech Coding Tool Set are summarized in Table 1.

MPEG-4 is based on tools each of which can be combined according to the user needs. HVXC consists of LSP (line spectral pair) VQ (vector quantization) tool and harmonic VQ tool. RPE (regular pulse excitation) tool, MPE (multipulse excitation) tool, and LSP VQ tool form CELP. RPE tool is allowed only for the wideband mode because of its simplicity at the expense of the quality. LSP VQ

Table 1
Specifications of MPEG-4 Natural Speech Coding Tools

HVXC		
Sampling frequency	8 kHz	
Bandwidth	300–3400 Hz	
Bit-rate (bit/s)	2000 and 4000	
Frame size	20 ms	
Delay	33.5–56 ms	
Features	Multi-bit-rate coding Bit-rate scalability	
CELP		
Sampling frequency	8 kHz	16 kHz
Bandwidth	300–3400 Hz	50–7000 Hz
Bit-rate (bit/s)	3850–12 200 (28 bit-rates)	10 900–23 800 (30 bit-rates)
Frame size	10–40 ms	10–20 ms
Delay	15–45 ms	15–26.75 ms
Features	Multi-bit-rate coding Bit-rate scalability Bandwidth scalability	

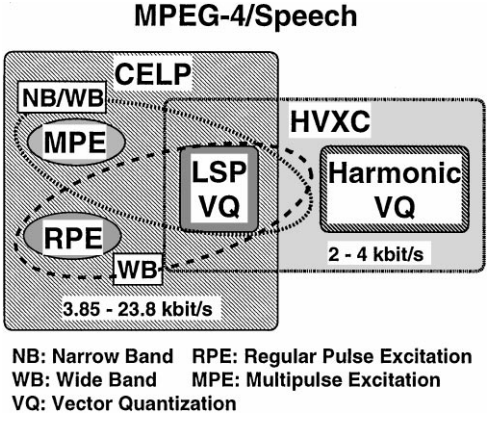


Fig. 6. MPEG-4 Natural Speech Coding Tool Set.

tool is common in both HVXC and CELP. MPEG-4 Natural Speech Coding Tools are illustrated in Fig. 6.

5.3. Functionalities of MPEG-4 Natural Speech Coding Tools

MPEG-4 Natural Speech Coding Tools are different from other existing speech coding standards such as ITU-T G.723.1 and G.729 in the following three new functionalities: multi-bit-rate coding,[1] bit-rate scalable coding, and bandwidth scalable coding. Actually, these new functionalities characterize MPEG-4 Natural Speech Coding Tools. It should be noted that the bandwidth scalability is available only for CELP.

5.3.1. Multi-bit-rate coding

Multi-bit-rate coding provides flexible bit-rate selection with the same coding algorithm. It has not been available and different codecs were needed for different bit-rates. In multi-bit-rate coding, a bit-rate is selected among multiple available bit-rates upon establishment of a connection between the communicating parties. The bit-rate for CELP may be selected with as small a step as 0.2 kbit/s. The frame length, the number of subframes per frame, and selection of the excitation codebook are modified for different bit-rates [17]. For HVXC, 2 or 4 kbit/s can be selected as the bit-rate.

In addition to multi-bit-rate coding, bit-rate control with a smaller step of the bit-rate is available for CELP by fine-rate control (FRC). In addition to multi-bit-rate coding, some additional bit-rates not available by multi-bit-rate coding are provided by FRC. The bit-rate may be deviated frame by frame from a specified bit-rate according to the input-signal characteristics. When the spectral envelope, approximated by the LP synthesis filter, has small variations in time, transmission of liner-prediction coefficients may be skipped once every two frames for a reduced average bit-rate [22].

Linear prediction coefficients in the current and the following frames are compared to decide if those in the following frame are to be transmitted or not. In the decoder, the missing LP coefficients in a frame are interpolated from those in the previous and the following frames. Therefore, FRC requires one-frame delay to make the data in the following frame available in the current frame.

5.3.2. Scalable coding

Bit-rate and bandwidth scalabilities are useful for multicast transmission. The bit-rate and the

[1] An arbitrary bit-rate may be selected with a 200 bit/s step by simply changing the parameter values.

bandwidth can be independently selected for each receiver by simply stripping off a part of the bit-stream. Scalabilities necessitate only a single encoder to transmit the same data to multiple points connected at different rates. Such a case can be found in connections between a cellular network with mobile terminals and a digital network with fixed multimedia terminals as well as in multipoint teleconferencing. The encoder generates a single common bit-stream by scalable coding for all the recipients instead of independent bit-streams at different bit-rates.

The scalable bit-stream has a layered structure with the core bit-stream and enhancement bit-streams. The bit-rate control is performed by adjusting the combination of the enhancement bit-streams depending on the specified bit-rate. The core bit-stream guarantees, at least, reconstruction of the original speech signal with a minimum speech quality. Additional enhancement bit-streams, which may be available depending on the network condition, will increase the quality of the decoded signal. HVXC and CELP may be used to generate the core bitstream when the enhancement bit-streams are generated by TwinVQ or AAC. They can also generate both the core and the enhancement bit-streams. Scalabilities in MPEG-4/CELP are depicted in Fig. 7.

Scalabilities include bit-rate scalability and bandwidth scalability. These scalabilities reduce signal distortion or achieve better speech quality with high-frequency components by adding enhancement bit-streams to the core bit-stream. These enhancement bit-streams contain detailed characteristics of the input signal or components in higher frequency bands. For example, the output of Decoder A in Fig. 7 is the minimum-quality signal decoded from the 6 kbit/s core bit-stream. The Decoder B output is a high-quality signal decoded from an 8 kbit/s bit-stream. Decoder C provides a higher-quality signal decoded from a 12 kbit/s bitstream. On the other hand, the Decoder D output has a wider bandwidth. This wideband signal is decoded from a 22 kbit/s bitstream. The high-frequency components of 10 kbit/s provides increased naturalness than Decoder C. Bandwidth scalability is provided only by the MPE tool.

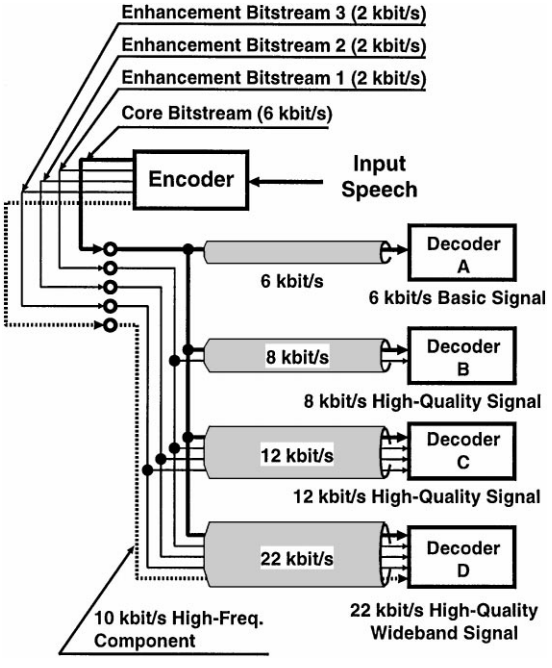


Fig. 7. Scalabilities in MPEG-4/CELP.

Table 2
Bandwidth scalable bit-streams

Core bit-stream (bit/s)	Enhancement bit-stream (bit/s)
3850–4650	9200, 10 400, 11 600, 12 400
4900–5500	9467, 10 667, 11 867, 12 667
5700–10 700	10 000, 11 200, 12 400, 13 200
11 000–12 200	11 600, 12 800, 14 000, 14 800

The unit bit-rate for the enhancement bit-streams in bit-rate scalability is 2 kbit/s for the narrowband and 4 kbit/s for the wideband. In case of bandwidth scalable coding, the unit bit-rate for the enhancement bit-streams depends on the total bit-rate and is summarized in Table 2.

5.4. Outline of the algorithms

5.4.1. HVXC

A basic blockdiagram of HVXC is depicted in Fig. 8. HVXC first performs LP analysis to find the LP coefficients. Quantized LP coefficients are

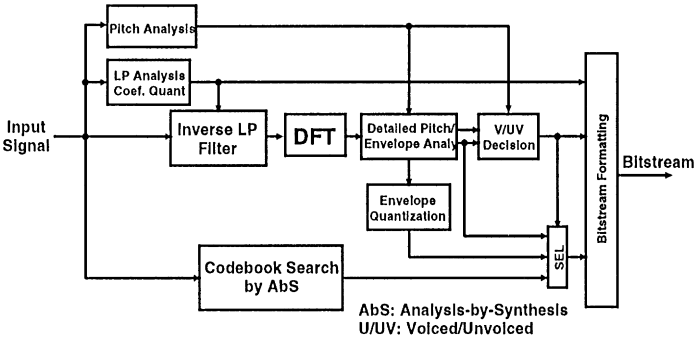


Fig. 8. HVXC.

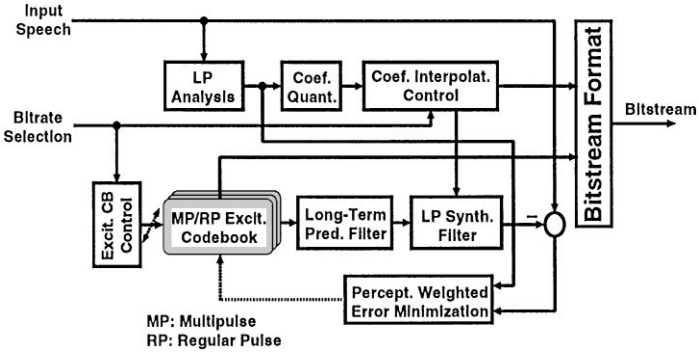


Fig. 9. CELP.

supplied to the inverse LP filter to find the prediction error. The prediction error is transformed into a frequency domain and the pitch and the envelope of the spectrum are analyzed. The envelope is quantized by weighted vector quantization in voiced sections. In unvoiced sections, closed-loop search of an excitation vector is carried out.

5.4.2. CELP

Fig. 9 shows a blockdiagram of CELP. The LP coefficients of the input signal are first analyzed and then quantized to be used in an LP synthesis filter driven by the output of the excitation codebooks. Encoding is performed in two steps. Long-term prediction coefficients are calculated in the first step. In the second step, a perceptually weighted error between the input signal and the output of the LP synthesis filter is minimized. This minimization is achieved by searching for an appropriate codevector for the excitation codebooks. Quantized coefficients, as well as indexes to the codevectors of the excitation codebooks and the long-term prediction coefficients, form the bit-stream. The LP coefficients are quantized by vector quantization and the excitation can be either MPE [19] or regular pulse excitation RPE [13].

MPE and RPE both model the excitation signal by multiple pulses, however, a difference exists in the degrees of freedom for pulse positions. MPE allows more freedom on the interpulse distance than RPE which has a fixed interpulse distance. Thanks to such a flexible interpulse distance, MPE achieves better coding quality than RPE [7]. On the other hand, RPE requires less computations than MPE by trading off its coding quality. Such a low computational requirement is useful in the wideband coding where the total computation should naturally be higher than in the narrowband coding. The excitation signal types of MPEG-4/CELP are summarized in Table 3.

Table 3
CELP excitation signal

Excitation	Bandwidth	Features
MPE	Narrow, wide	Quality, scalability
RPE	Wide	Complexity

5.5. MPEG-4/CELP with MPE

MPEG-4/CELP with MPE is the most complete combination of the tools in MPEG-4 Natural Speech Coding Tools. It provides all the three new functionalities. Therefore, it is useful to explain MPEG-4/CELP with MPE in more detail to show how these functionalities are realized in the algorithm.

A blockdiagram of the encoder of MPEG-4/CELP with MPE is depicted in Fig. 10. It consists of three modules; a CELP core encoder, a bit-rate scalable (BRS) tool, and a bandwidth extension (BWE) tool. The CELP core encoder provides the basic coding functions which have been explained with Fig. 9 in Section 5.4.2. The BRS tool is used to provide the bit-rate scalability. The residual of the narrowband signal, mode information, LP coefficients, quantized LSP coefficients, and multipulse excitation signal are transferred from the core encoder to the BRS tool as the input signals. The BWE tool is used for the bandwidth scalability.

Quantized LSP coefficients and the pitch delay indexes as well as the wideband speech to be encoded are supplied from the core encoder to the BWE tool.

In addition to these input signals, the narrowband multipulse excitation is needed in the BWE tool. This excitation is supplied from either the BRS tool when the bit-rate scalability is implemented, or from the core encoder. When the bandwidth scalability is provided, a downsampled narrowband signal is supplied to the core encoder. Because of this downsampling operation, an additional 5-ms look-ahead of the input signal is necessary for wideband signals.

5.5.1. CELP core encoder

Fig. 11 depicts a blockdiagram of the CELP core encoder. It performs LP analysis and pitch analysis on the input speech signal. The obtained LP coefficients, the pitch lag (phase or delay), the pitch and MPE gains, and the excitation signal are encoded as well as mode information. The LP coefficients in the LSP domain are encoded frame by frame by predictive VQ. The pitch lag is encoded subframe by subframe by adaptive codebooks. The MPE is modeled by multiple pulses whose positions and polarities (± 1) are encoded. The pitch and the MPE gains are normalized by an average subframe power followed by multimode encoding [19]. The average subframe power is scalar-quantized in each frame.

5.5.1.1. LSP quantization. A two-stage partial prediction and multistage vector quantization (PPM-VQ) [21] is employed for LSP quantization. This

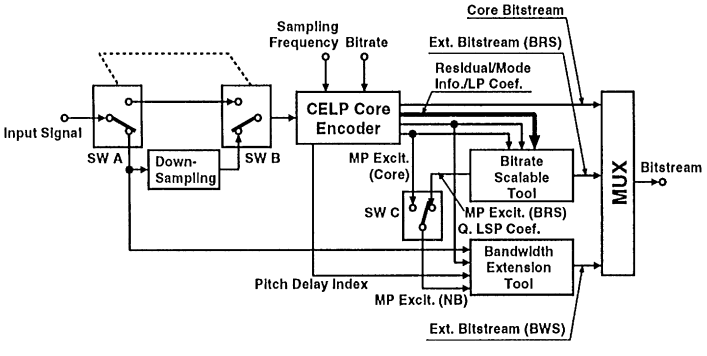


Fig. 10. MPEG-4/CELP with MPE.

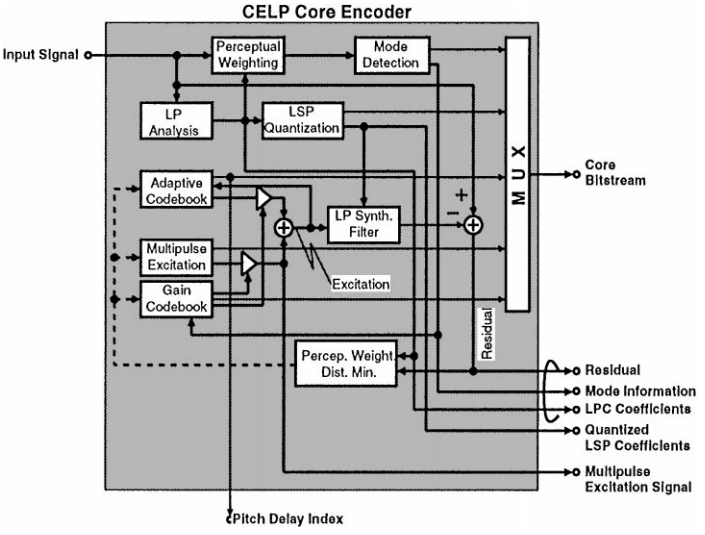


Fig. 11. CELP core encoder.

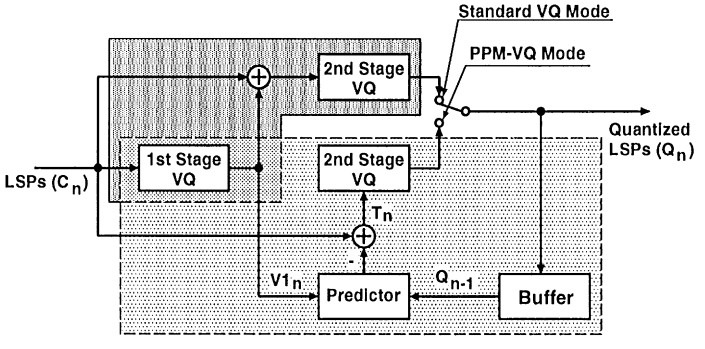


Fig. 12. Partial prediction and multistage vector quantization (PPM-VQ).

quantizer, as shown in Fig. 12, operates either in the standard VQ mode or in the PPM-VQ mode which utilizes interframe prediction, depending on the quantization errors. The standard VQ mode operates as a common two-stage VQ which quantizes the error of the first stage in the second stage. On the contrary, in the PPM-VQ mode, the difference T_n between the input LSP C_n and its predicted output is quantized as

$$T_n = C_n - (\beta_p Q_{n-1} + (1.0 - \beta_p) V1_n). \quad (1)$$

The second term of (1) is the predicted output which is obtained from the quantized output $V1_n$ of the first stage and the quantized LSP Q_{n-1} in the previous frame. β_p stands for the prediction coefficient and is set to 0.5 in MPEG-4/CELP.

PPM-VQ provides good coding quality in both stationary and nonstationary speech sections by appropriately selecting the predictive VQ or the standard VQ. Transmission-error propagation dies out quickly because prediction is employed only in the second stage. The number of LSPs is 10 for the narrowband and 20 for the wideband case. Because the wideband mode has twice as many parameters, two narrowband quantizers connected in parallel are used; one for the first 10 parameters and the

other for the rest, respectively. The number of bits used for LSP quantization is 22 for the narrowband and 46 for the wideband (25 for the first 10 coefficients and 21 for the rest). The codebook has 1120 words for the narrowband and 2560 words for the wideband.

5.5.1.2. Multipulse excitation. The multipulse excitation μ_n has L pulses as in

$$\mu_n = \sum_{i=1}^{L} s_{m_i} \delta_{n-m_i}, \quad n = 0, \ldots, N-1, \tag{2}$$

where N stands for the subframe size, and m_i and s_{m_i} are the position and the magnitude of the ith pulse, respectively. The pulse position is selected from M_i candidates which are defined by the Algebraic code [1,14] for each pulse. The pulse magnitude is represented only by its polarity for bit reduction. Such a simplified excitation model contributes to reduced computations compared with conventional CELP codebooks at a low bit-rate with a small number of pulses. On the other hand, reduction of computations is necessary for a high bit-rate with more available pulses. For example, MPE encoding by tree search [16] provides easy bit-rate control by adjusting the number of pulses. Efficient coding techniques by combination search of the pulse position and polarity and by VQ of the pulse polarity [19] may also be applied. These additional techniques help us avoid reduced quality and heuristic parameter setting caused by well-known preselection techniques and focused search [14] for the pulse position.

5.5.2. Bit-rate scalable (BRS) tool

A blockdiagram of the BRS tool [18] is shown in Fig. 13. The actual signal to be encoded in the BRS tool is the residual, which is defined as the difference between the input signal and the output of the LP synthesis filter (local decode signal), supplied from the core encoder. This combination of the core encoder and the BRS tool can be considered as multistage encoding of the MPE. However, there is no feedback path for the residual in the BRS tool connected to the MPE in the core encoder. The excitation signal in the BRS tool has no influence on the adaptive codebook in the core encoder. This guarantees that the adaptive codebook in the core decoder at any site is identical to that in the encoder (in terms of the codewords), which leads to the minimum quality degradation for the frame-by-frame bit-rate change. The BRS tool adaptively controls the pulse positions so that none of them coincides with a position used in the core encoder. This adaptive pulse position control contributes to more efficient multistage encoding.

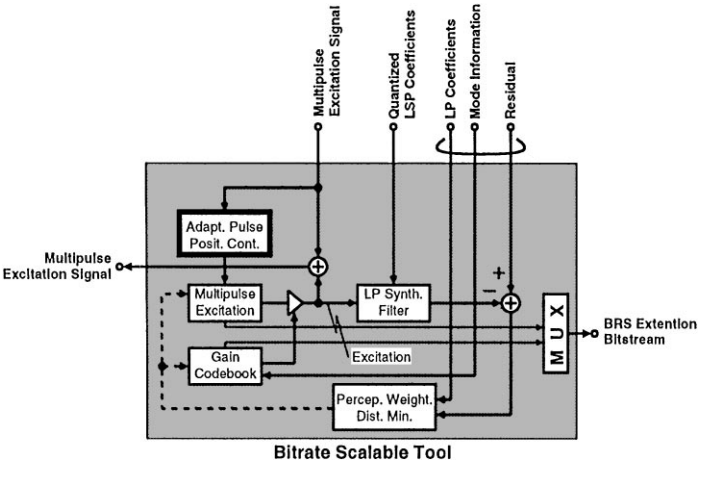


Fig. 13. Bit-rate scalable (BRS) tool.

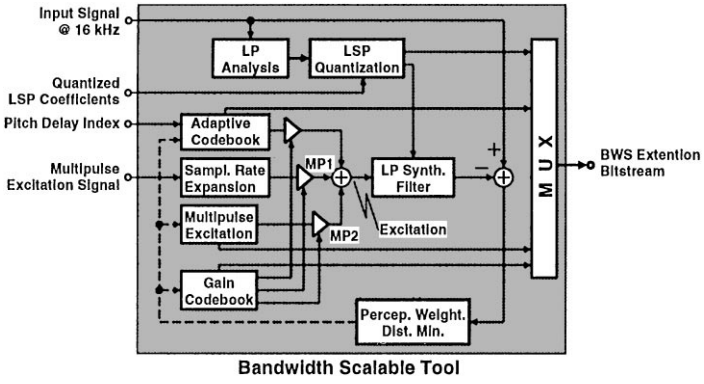


Fig. 14. Bandwidth extension (BWE) tool.

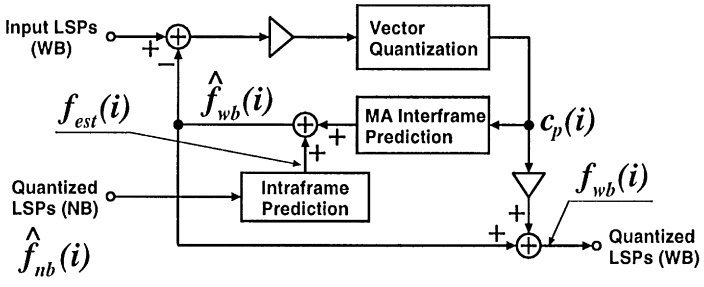


Fig. 15. LSP quantization in the BWE tool.

5.5.3. Bandwidth extension (BWE) tool

Fig. 14 exhibits a blockdiagram of bandwidth extension (BWE) tool [15]. The BWE tool is also a CELP-based encoder and encodes the frequency components which are not processed by the narrowband core encoder as well as a fraction of the narrowband components which have not been encoded. Quantized LSP coefficients and excitation signals of the narrowband components are supplied from the core encoder, in addition to the pitch delay index.

5.5.3.1. LSP quantization. A blockdiagram of LSP quantization in the BWE tool is shown in Fig. 15. Predicted wideband LSPs are subtracted from the input LSPs and the residuals are vector-quantized. The indexes to the codevectors of the codebook are incorporated in the output bit-stream. The vector-quantized residual is added to the predicted wideband LSPs to reconstruct quantized LSPs. These quantized LSPs are supplied to the LP synthesis filter. The predicted wideband LSPs, $\hat{f}_{\mathrm{wb}}(i)$ for $i = 1, \ldots, N_{\mathrm{wb}}$, are constructed by adding estimated wideband LSPs $f_{\mathrm{est}}(i)$ for $i = 1, \ldots, N_{\mathrm{wb}}$ to interframe prediction of the quantized residuals based on a moving-average as in (3).

$$\hat{f}_{\mathrm{wb}}(i) = \sum_{p=0}^{P} a_p(i) c_p(i) + f_{\mathrm{est}}(i), \quad i = 1, \ldots, N_{\mathrm{wb}}, \qquad (3)$$

where $a_p(i)$ is the interframe prediction coefficient and P is the prediction order. $c_p(i)$ is the quantized prediction residual in the pth previous frame.

Estimated wideband LSPs are obtained by scaling the quantized narrowband LSPs to the wideband as shown in the following equation with

a scaling factor $b(i)$:

$$f_{\mathrm{est}}(i) = \begin{cases} b(i)\hat{f}_{\mathrm{nb}}(i) & \text{for } i = 1, \ldots, N_{\mathrm{nb}}, \\ 0.0 & \text{for } i = N_{\mathrm{nb}} + 1, \ldots, N_{\mathrm{wb}}, \end{cases} \quad (4)$$

$\hat{f}_{\mathrm{nb}}(i)$ represents the ith quantized narrowband LSP.

This algorithm provides better quantization precision for low-order LSPs ($f_{\mathrm{wb}}(i)$, $i = 1, \ldots, N_{\mathrm{nb}}$) as well as for high-order LSPs ($f_{\mathrm{wb}}(i)$, $i = N_{\mathrm{nb}} + 1, \ldots, N_{\mathrm{wb}}$). This is because the residual LSPs to be vector-quantized contain narrowband LSP residuals which have not been taken care of in the narrowband core encoder.

5.5.3.2. Multipulse excitation. The excitation signal in the bandwidth extension tool is represented by an adaptive codebook, two MPE signals, and their gains as shown in Fig. 14. The pitch delay of the adaptive codebook is searched for from the vicinity of its estimation obtained from the narrowband pitch-delay. One of the two MPE signals (MP1) is an upsampled version of the narrowband MPE signal and the other (MP2) is an exclusive MPE signal in the bandwidth extension tool. The adaptive codebook and the gains for MP2 are vector-quantized and the gains for MP1 are scalar-quantized. These quantizations are performed to minimize the perceptually weighted error.

5.6. Coding quality of MPEG-4 Natural Speech Coding Tools

Coding quality of CELP [7] is depicted in Fig. 16(a) and (b). The narrowband mode with MPE, which supports multiple bit-rates with a single algorithm, exhibits comparable quality to those of other standards (ITU-T G.723.1, G.729, ETSI GSM-EFR) optimized at a single bit-rate. The wideband mode with MPE at 17.9 kbit/s provides comparable quality to those of ITU-T G.722 at 56 kbit/s and MPEG-2 Layer III at 24 kbit/s.

Speech quality by scalable coding naturally is degraded compared to nonscalable coding because of split loss of the bit-stream. However, MPEG-4 CELP is successful in minimizing the difference in speech quality between scalable and nonscalable coding thanks to its algorithmic features.

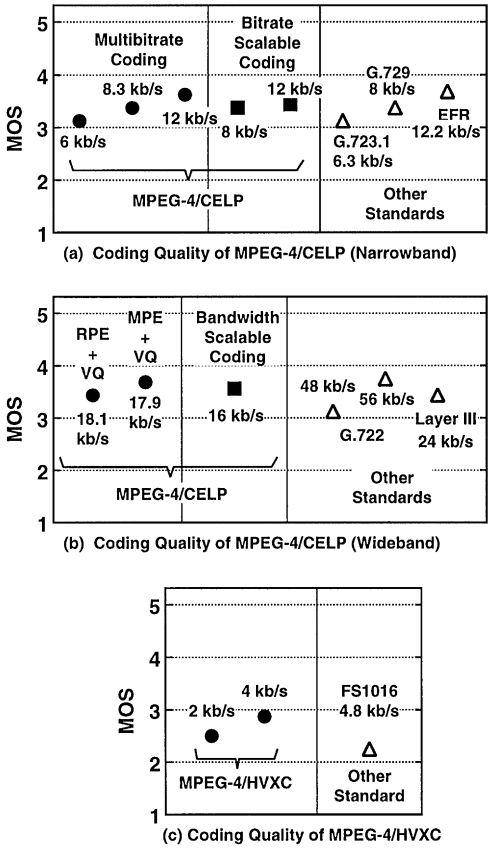


Fig. 16. Coding quality of MPEG-4 Natural Speech Coding Tools.

Fig. 16(c) exhibits the coding quality of HVXC [7]. MPEG-4/HVXC clearly outperforms the reference, FS1016 [3] at 4.8 kbit/s.[2]

6. Scalability

(Bit-stream) scalability is the ability of an audio codec to support an ordered set of bit-streams which can produce a reconstructed sequence. Moreover, the codec can output useful audio when certain subsets of the bit-stream are decoded. The

[2] FS1016 is a US Department of Defence (DoD) standard that is most commonly used as a reference at a bit-rate lower than 5 kbit/s.

minimum subset that can be decoded is called the base layer. The remaining bit-streams in the set are called enhancement or extension layers. Depending on the size of the extension layers we talk about large step or small step (granularity) scalability. Small step scalability denotes enhancement layers of around 1 kbit/s (or smaller). Typical data rates for the extension layers in a large step scalable system are 16 kbit/s or more. Scalability in MPEG-4 natural audio largely relies on difference encoding, either in time domain or, as in the case of AAC layers, of the spectral lines (frequency domain).

6.1. Comparison to simulcast

A trivial way to implement bit-stream scalability is the simulcast of several bit-streams at different bit-rates. Especially in the case of just two layers of scalability, this solution has to be checked against a more complex "real" scalable system. Depending on the size of the enhancement layers, a scalable system has to take a hit in compression efficiency compared to a similar non-scalable system. Depending on the algorithm, this cost (in terms of bit-rate for equivalent quality) can vary widely. For the scalable systems defined in MPEG-4 natural audio, the cost has been estimated in several verification tests. In each of the cases, the scalable system performed better than the equivalent simulcast system. In the optimum case it may be found that the scalable system is improved over the equivalent non-scalable system at the same bit-rate. This is expected to happen only for certain combinations and signal classes. An example for this effect is the combination of a speech core coder based on CELP (building on a model of the human vocal tract to enhance the speech quality) and enhancement layers based on AAC (to get higher quality especially for non-speech signals and at higher bit-rates). This combination may perform better than AAC for speech signals alone. While the effect has been demonstrated during the core experiment process, it did not show up in the verification test results.

Scalability is at the heart of the new MPEG-4 audio functionalities. Some sort of scalability has been built into all of the MPEG-4 natural audio coding algorithms.

6.2. Types of scalability in MPEG-4 natural audio

MPEG-4 natural audio allows for a large number of codec combinations for scalability. The combinations for the speech coders are described in the paragraphs explaining MPEG-4 CELP and HVXC. The following list contains the main combinations for MPEG-4 general audio (GA):

- AAC layers only,
- Narrow-band CELP base layer plus AAC,
- TwinVQ base layer plus AAC.

Depending on the application, either of these possibilities can provide optimum performance. In all cases where good speech quality at low bit-rates is a requirement for the case of reception of the core layer only (like for example in a digital broadcasting system using hierarchical channel coding), the speech codec base layer is preferred. If, on the other hand, music should be of reasonable quality for a very low bit-rate core layer (for example for Internet streaming of music using scalability), the TwinVQ base layer provides the best quality. If the base layer is allowed to work at somewhat higher bit-rates (like 16 bit/s or more), a system built from AAC layers only can deliver the best overall performance.

6.3. Block length considerations

In the case of combining speech coders and General Audio coding, special consideration has to be given to the frame length of the underlying coding algorithms. This is trivial in the case of different AAC layers at the same sampling frequency. For the speech coders in MPEG-4 natural audio, the frame length is a multiple of 10 ms which does not match the frame lengths normally used in MPEG-4 GA. To accommodate these different frame length, two modifications have been done to the scalable system:

AAC modified block length. A modified AAC works at a basic block length of 960 samples (instead of the usual 1024). This translates to a block length of 20 ms at 48 kHz sampling frequency. At the other main sampling frequencies for scalable MPEG-4 AAC, the basic block length of the AAC enhancement layers is again a multiple of 10 ms.

Super frame structure. To keep a frame a single decodable instance of audio data, several data blocks may be combined into one super-frame. For example, at a sampling frequency of 16 kHz and a core block length for a CELP core of 20 ms, three CELP blocks and one block of AAC enhancement layers are combined into one super-frame.

6.4. Mono–stereo scalability

At low bit-rates, mono transmission is often preferred to stereo at the same total bit-rates. Most listeners evaluate the degradation due to the overhead of stereo transmission to be more annoying than the loss of stereo. For higher bit-rates, stereo transmission is virtually a requirement today. Therefore, stereo enhancement layers can be added as enhancement layers to both mono and stereo lower layers.

6.5. Overview of scalability modes in MPEG-4 natural audio

Table 4 lists the possibilities for scalability layers within MPEG-4 natural audio. All narrowband CELP (mono), TwinVQ (mono), TwinVQ (stereo), AAC (mono) and AAC (stereo) can be used as core layers. Enhancement layers can be of the types NB CELP mono (on top of CELP only), TwinVQ mono (on top of TwinVQ mono only), TwinVQ stereo (on top of TwinVQ stereo only), AAC mono (on top of NB CELP, TwinVQ mono or AAC mono) or AAC stereo (on top of any of the other codecs).

6.6. Frequency-selective switch (FSS) module

Not in all cases the difference signal between the output of a lower layer and the original (frequency domain) signal is the best input to code an enhancement layer. If, for instance a scalable coder using a CELP core coder would be used to encode musical material, the output of the CELP coder may be able to help the enhancement layers in terms of getting an easier signal to encode. To enable more flexible coding of enhancement layers, a frequency selective switch (FSS) module has been introduced. It basically consists of a bank of switches operating independently on a scalefactor band basis. For each scalefactor band, one of two inputs into the system can be selected.

6.7. Upsampling filter tool

For scalability spanning a wider range of bitrates (from speech quality to CD quality), it is not recommended to run the core coders at the same sampling frequency as the enhancement layer coders. To accommodate this requirement, an upsampling filter tool has been defined. It uses the MDCT (very similar to the IMDCT already present in the AAC decoder) algorithm to perform the filtering. A number of zeroes is inserted into the time domain waveform and used as the input to the MDCT. The output values can then directly combined with MDCT values from a higher sampling frequency filter bank. The prototype filter in this case is the MDCT window function and is the same as used in the AAC IMDCT.

Table 4
Overview of scalability modes

Layer N	NB CELP mono	TwinVQ mono	TwinVQ stereo	AAC mono	AAC stereo
Narrowband CELP mono	X			X	X
TwinVQ mono		X			X
TwinVQ stereo			X		X
AAC mono				X	X
AAC stereo					X

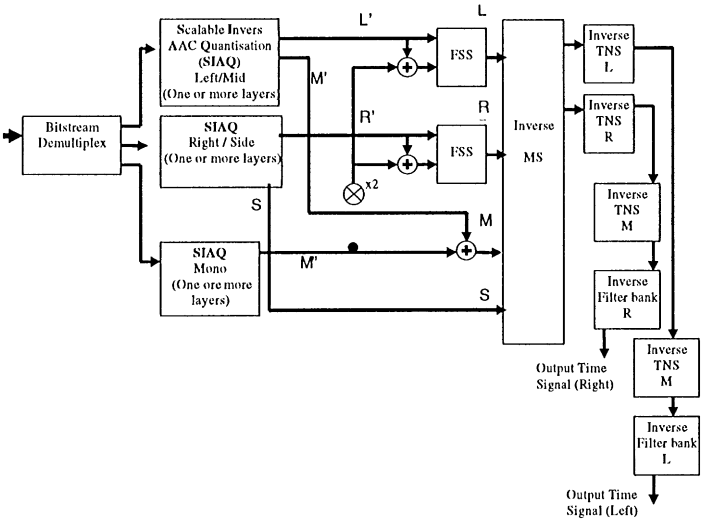


Fig. 17. Scalability example.

6.8. A scalability example

The following example illustrates the combined use of a number of the tools implementing scalability. Fig. 17 shows the decoding of a non-GA (i.e. CELP) mono plus AAC stereo combination and can be found in ISO/IEC International Standard 14496-3 (MPEG-4 audio). A mono CELP is combined with a single stereo AAC layer. Temporal noise shaping (TNS) is applied to the MDCT coefficients calculated from the upsampled CELP decoder output. Three FSS (frequency selective switch) modules either combine the upsampled and TNS processed output signal from the core coder with the AAC decoded spectral data or use just the AAC spectral data. Full M/S processing is possible in this combination to enhance the stereo coding efficiency. A core only decoder just uses the CELP data and applies normal CELP postfiltering. The CELP output for higher-quality decoding is not postfiltered. Depending on the number of scalability layers, much more complex structures need to be built for MPEG-4 audio scalable decoding.

7. Audio profiles and levels

In order to maximize interoperability, only a small number of profiles have been defined for MPEG-4 Audio. Given the rather large number of coding tools and object types, this leads to the inclusion of a relatively large number of audio object types even in the more simple profiles. Some of the audio profiles (e.g. main profile) contain both natural and structured audio object types. The following table lists only the audio profiles containing natural audio object types:

Speech coding profile. This profile contains the CELP and HVXC object types as well as an interface to text-to-speech (TTS).

Scalable profile. This profile contains the lower complexity AAC object types (LC, LTP), both in MPEG-2 (IS 13818-7) style and using the syntax which enables scalability (MPEG-4 style). In addition, the TwinVQ and speech coding object types and all the tools for scalable audio are part of this profile. It is expected, that most early applications will use this profile.

Main profile. This is the "do-it-all" profile of MPEG-4 natural and structured audio. It contains all the MPEG-4 audio coding tools.

A hierarchical organisation of the profiles supports the "design for interoperability": The speech coding profile is contained in all the other profiles containing natural audio coding tools, the scalable audio profile is contained in the main profile.

Table 5 lists all the tools of MPEG-4 natural audio and their use in the different audio objects types.

Table 5
Usage of audio object types

Tools / Audio object types	13818-7 main	13818-7 LC	13818-7 SSR	PNS	LTP	TLSS	TwinVQ	CELP	HVXC	GA bit-stream syntax type	Hierarchy
AAC main	X			X						ISO\IEC 13818-7 Style	Contains AAC LC
AAC LC		X		X						ISO\IEC 13818-7 Style	
AAC SSR			X	X						ISO\IEC 13818-7 Style	
AAC LTP		X		X	X					ISO\IEC 13818-7 Style	Contains AAC LC
AAC scalable		X		X	X	X				Scalable	
TwinVQ					X		X			Scalable	
CELP								X			
HVXC									X		

Table 6
Decoder complexity

Object type	Parameter f_s (kHz)	PCU (MOPS)	RCU (kWords)
AAC Main[a]	48	5	5
AAC LC[a]	48	3	3
AAC SSR[a]	48	4	3
LTP[a]	48	4	4
AAC Scalable[a,b]	48	5	4
TwinVQ[a]	24	2	3
CELP	8	1	1
CELP	16	2	1
CELP	8/16	3	1
HVXC	8	2	1

Definitions: f_s = sampling frequency.
Notes:
[a] PCU proportional to sampling frequency.
[b] Includes core decoder.

7.1. Levels for the MPEG-4 audio scalable profile

The large number of possibilities to combine different audio object types makes the traditional way of defining levels according to the channel count, sampling frequency, etc. very difficult. In order to enable decoder implementers to conform with a certain level definition and still retain the possibility to combine different audio object types, complexity units have been defined and are used to calculate necessary decoder capabilities. For each audio object type, the decoder complexity (for a given sampling rate and channel count) was estimated in PCUs (computing complexity counted as millions of operations per second needed) and RCUs (memory complexity counted in kWords buffer requirements). Of course these complexity numbers depend a lot on the architecture of a decoder, whether realized on a dedicated DSP or a general purpose computing architecture. Table 6 lists the estimates of decoders for different object types as submitted to the MPEG audio group:

The level of a scalable profile decoder can now be determined by PCU and RCU numbers in addition to the number of channels and sampling frequencies. Four levels have been defined. They are:

Level 1. One mono object of up to 24 kHz sampling frequency, all object types.

Level 2. One stereo or two mono objects of up to 24 kHz sampling frequency.

Level 3. One stereo or two mono objects of up to 48 kHz sampling frequency.

Level 4. One 5.1 channel object or a flexible configuration of objects up to 48 kHz sampling frequency and a PCU up to 30 and RCU up to 19.

Acknowledgements

The authors would like to thank Jürgen Koller from Fraunhofer Institut Integrierte Schaltungen for the support while writing this paper as well as Dr. Kazunori Ozawa, Dr. Masahiro Serizawa, and Mr. Toshiyuki Nomura of C&C Media Research Laboratories, NEC Corporation, for their valuable comments and discussions. Thanks to Bodo Teichmann and Bernhard Grill for supplying figures. Part of the work at Fraunhofer IIS was supported by the European Commission (ACTS MoMuSys) and the Bavarian Ministry for Economy, Transportation and Technology. MPEG-4 natural audio is the joint effort of numerous people who worked hard to make the vision of a flexible multimedia coding system a reality. We want to mention the work of Peter Schreiner, the chair of the audio subgroup in MPEG, and David Meares, the audio secretary, and all the others who worked together to create MPEG-4 audio.

References

[1] J-P. Adoul, P. Mabilleau, M. Delprat, S. Morissette, Fast CELP coding based on algebraic codes, in: Proceedings of ICASSP'87, April 1987, pp. 1957–1960.

[2] M. Bosi, K. Brandenburg, S. Quackenbusch, K. Akagiri, H. Fuchs, J. Herre, L. Fielder, M. Dietz, Y. Oikwa, G. Davidson, ISO/IEC MPEG-2 Advanced Audio Coding, presented at the 101st Convention of the Audio Engineering Society, J. Audio Eng. Soc. (Abstracts) 44 (December 1996) 1174, preprint 4382.

[3] J.P. Campbell, T.E. Tremain, V.C. Welch, The DOD 4.8KBPS standard (Proposed Federal Standard 1016), in: Advances in Speech Coding, Kluwer Academic Publishers, Boston, 1991, pp. 121–133.

[4] A. Gersho, Advances in speech and audio compression, Proc. IEEE 82 (6) (June 1994) 900–918.

[5] J. Herre, J.D. Johnston, Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS), Presented at the 101st Convention of the Audio Engineering Society, preprint 4384.

[6] J. Herre, D. Schulz, Extending the MPEG-4 AAC codec by perceptual noise substitution, Presented at the 104st Convention of the Audio Engineering Society, preprint 4720.

[7] ISO/IEC JTC1 SC29/WG11, Report on the MPEG-4 Speech Codec Verification Test, ISO/IEC JTC1/SC29/WG11 N2424, October 1998.

[8] ISO/IEC JTC1 SC29/WG11, ISO/IEC FDIS 14496-3 Subparts 1, 2, 3, Coding of Audio–Visual Objects – Part 3: Audio, ISO/IEC JTC1 SC29/WG11 N2503, October 1998.

[9] ISO/IEC JTC1 SC29/WG11 N2725, Overview of the MPEG-4 Standard, Seoul, 1999.

[10] N. Iwakami, T. Moriya, Transform domain weighted interleave vector quantization (Twin VQ), Presented at the 101st Convention of the Audio Engineering Society, preprint 4377.

[11] N. Iwakami, T. Moriya, The integrated filterbank based scalable MPEG-4 audio coder, Presented at the 105th Convention of the Audio Engineering Society, preprint 4810.

[12] N.S. Jayant, High-quality coding of telephone speech and wideband audio, IEEE Commun. Mag. (January 1990) 10–20.

[13] P. Kroon, E.F. Deprettere, R.J. Sluyter, Regular-pulse excitation – A novel approach to effective and efficient multipulse coding of speech, IEEE Trans. SP ASSP-34 (5) (October 1986) 1054–1063.

[14] C. Laflamme, J.-P. Adoul, R. Salami, S. Morissette, P. Mabilleau, 16 kbps wideband speech coding technique based on algebraic CELP, in: Proceedings of ICASSP'91, May 1991, pp. 13–16.

[15] T. Nomura, M. Iwadare, M. Serizawa, K. Ozawa, A bitrate and bandwidth scalable CELP coder, in: Proceedings of ICASSP 98, May 1998, Vol. I, pp. 341–344.

[16] T. Nomura, K. Ozawa, M. Serizawa, Efficient pulse excitation search methods in CELP, Nat. Conf. Proc. Acoust. Soc. J. 2-P-5 (March 1996) 311–312.

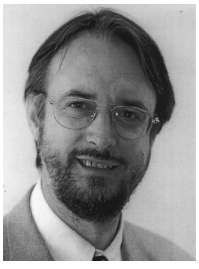
[17] T. Nomura, M. Serizawa, K. Ozawa, An MP-CELP speech coding algorithm with bit rate control, in: Proceedings of the IEICE General Conference, SD-5-3, March 1997, pp. 348–349.

[18] T. Nomura, M. Serizawa, K. Ozawa, An embedded MP-CELP speech coding algorithm using adaptive pulse-position control, in: Proceedings of IEICE Society Conference, September 1997, Vol. D, pp. D-14–10.

[19] K. Ozawa, M. Serizawa, T. Miyanao, T. Nomura, M. Ikekawa, S. Taumi, M-LCELP speech coding at 4 kb/s with multi-mode and multi-codebook, IEICE Trans. E77-B (9) (September 1994) 1114–1120.

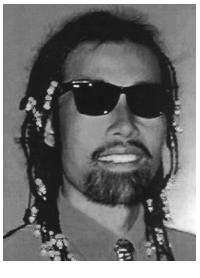
[20] D. Schulz, Improving audio codecs by noise substitution, J. AES 44 (7/8) (July/August 1996) 593–598.

[21] N. Tanaka, T. Morii, K. Yoshida, K. Honma, A multimode variable rate speech coder for CDMA cellular systems, in: Proceedings of IEEE Vehicular Technology Conference 96, April 1996, pp. 198–202.
[22] R. Taori, R.J. Sluijter, A.J. Gerrits, On scalability in CELP coding systems, in: Proceedings of IEEE Speech Coding Workshop 97, September 1997, pp. 67–68.



Karlheinz Brandenburg was born in Erlangen, Germany in 1954. He received M.S. (Diplom) degrees in Electrical Engineering in 1980 and in Mathematics in 1982 from Erlangen University. In 1989 he earned his Ph.D. in Electrical Engineering, also from Erlangen University, for work on digital audio coding and perceptual measurement techniques. The techniques described in his thesis form the basis for MPEG-1/2 Audio Layer-3, MPEG-2 Advanced Audio Coding (AAC) and most other modern audio compression schemes. From 1989 to 1990 he was with AT&T Bell Laboratories in Murray Hill, NJ, USA. He worked on the ASPEC perceptual coding technique and on the definition of the ISO/IEC MPEG/Audio Layer-3 system. In 1990 he returned to Erlangen University to continue the research on audio coding and to teach a course on digital audio technology. Since 1993 he is department head at the Fraunhofer Institute für Integrierte Schaltungen (FhG-IIS) in Erlangen, Germany. He has presented numerous papers at AES conventions and IEEE conferences. Together with Mark Kahrs, he edited the book "Applications of Digital Signal Processing to Audio and Acoustics". In 1994 he received the AES Fellowship Award for his work on perceptual audio coding and psychoacoustics. In 1998 he received the AES silver medal award for "sustained innovation and leadership in the development of the art and science of perceptual encoding". Dr. Brandenburg is a member of the technical committee on Audio and Electroacoustics of the IEEE Signal Processing Society. He has worked within the MPEG-Audio committee since its beginnings in 1988. He served as editor of MPEG-1 Audio, MPEG-2 Audio and as adhoc chair for a number of adhoc groups during the development of MPEG-2 Advanced Audio Coding and MPEG-4 Audio. From 1995 on, under his direction Fraunhofer IIS developed copyright protection technology including secure envelope techniques (MMP, Multimedia Protection Protocol) and watermarking. Dr. Brandenburg has been granted 24 patents and has several more pending.

Oliver Kunz was born in 1968 in Hagen, Germany. He received his M.S. (Diplom) from Bochum University in 1995. His master thesis at the chair of Prof. Blauert was on a realtime model of binaural localisation of sound sources. He joined the department Audio/Multimedia at the Fraunhofer Institut für Integrierte Schaltungen (FhG-IIS) in 1995. He worked on the implementation and optimisation of a high-quality realtime MPEG Layer-3 encoder for the digital radio system WorldSpace and actively contributed to the standardisation of MPEG-2 AAC. Since January 1998 he is head of the Audio Coding group at FhG-IIS. Responsibilities of this group range from quality optimisation of state-of-the-art coding schemes and contributions to international standardisation bodies to audio broadcast system related issues.



Akihiko Sugiyama received the B. Eng., M. Eng., and Dr. Eng. degrees in electrical engineering from Tokyo Metropolitan University, Tokyo, Japan, in 1979, 1981, and 1998, respectively. He joined NEC Corporation, Kawasaki, Japan, in 1981 and has been engaged in research on signal processor applications to transmission terminals, subscriber loop transmission systems, adaptive filter applications, and high-fidelity audio coding. In the 1987 academic year, he was on leave at the Faculty of Engineering and Computer Science, Concordia University, Montreal, P.Q., Canada, as a Visiting Scientist. From 1989 to

1994, he was involved in the activities of the Audio Subgroup, ISO/IEC JTC1/SC29/WG11 (known as MPEG/Audio) for international standardization of high-quality audio data compression as a member of the Japanese delegation. His current interests lie in the area of signal processing and circuit theory. Dr. Sugiyama is a member of the *Institute of Electrical and Electronic Engineers* (IEEE) and the *Institute of Electronics, Information and Communication Engineers* (IEICE) of Japan. He served as an associate editor for the IEEE Transactions on Signal Processing from 1994 to 1996. He is also a member of the Technical Committee for Audio and Electroacoustics, IEEE Signal Processing Society. He is currently serving as an associate editor for the *Transactions of the IEICE* on Fundamentals of Electronics, Communications and Computer Sciences. He received the 1988 Shinohara Memorial Academic Encouragement Award from IEICE. He is a coauthor of *International Standards for Multimedia Coding* (Yokohama, Japan: Maruzen, 1991), *MPEG/International Standards for Multimedia Coding* (Tokyo, Japan: Ohmusha, 1996), *Digital Broadcasting* (Tokyo, Japan: Ohmusha, 1996), and *Digital Signal Processing for Multimedia Systems* (New York: Marcel Dekker, Inc., 1999). Dr. Sugiyama is the inventor of 50 registered patents in the US, Japan, Canada, Australia, and European Patent Committee (EPC), in the field of signal processing and communications.