Current Status of the MPEG-4 Audio Verification Model Development

Bernd Edler

Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung

University of Hannover, Germany

Abstract

After a short survey of the goals of the current standardization activities within MPEG-4 Audio, an introduction to the concept of the verification model (VM) based codec development is given and the current status of the VM development is described. The results of listening tests and core experiments are addressed briefly in order to allow a comparison of the expected codec efficiency with that of existing standards.

1 Introduction

The goal of the MPEG-4 Audio activities within the ISO/IEC MPEG working group is to develop a standard for very low bit rate audio coding suitable for various applications in the fields of multimedia storage and communication. Since an object based representation is desired which allows content based interactivity, in addition to pure coding efficiency with bit rates ranging from 2 kbit/s to 48 kbit/s per channel other functionalities are taken into account. One important functionality is scalability, i.e. the possibility to decode lower rate subsets of a higher rate data stream. Further functionalities include speed change and pitch change, i.e. the possibility for independent manipulation of the speed and/or pitch during sound reproduction. Low delay and low complexity modes are also envisaged as well as robustness in the case of channel errors during transmission. Profiles will be defined which contain detailed specifications for key applications. The status of an International Standard shall be reached by November 1998.

This paper gives an introduction in the current status of the activities which are reflected by the so-called Verification Model (VM). In Section 2 the concept of VM based codec development is briefly described. Section 3 summarizes the conditions and the results of the listening test which was carried out to assist the definition of an initial VM. The structure of the resulting VM is presented in Section 4. Some indications on possible core experiments which allow codec developers to prove the merits of proposals for improvements are given in Section 5. Conclusions are drawn in Section 6 which also gives a short outlook on future extensions of the VM.

2 Concept of Verification Model based Codec Development

In order to manage the task of considering various combinations of possible coding schemes, types of input signals and functionalities, the selection of key components and the optimization is based on a verification model (VM) [1]. VM-based codec development already has been carried out successfully within the MPEG video standardization and in the development of the MPEG-2 non backward compatible (NBC) multichannel audio codec [2]. The VM approach is based on

the definition of a coding scheme which is decomposed into modules with appropriately defined interfaces. This procedure allows to carry out so-called core experiments in which the individual modules can be optimized as independently as possible by comparing the performance of proposals for improved modules or even for replacements against the current VM. If a core experiment is successful, i.e. it proves the advantages of a newly proposed module, this new module is incorporated in the VM.

3 Listening Tests

For the construction of an initial VM a pre-selection of promising approaches was necessary, and therefore a first subjective test was carried out. For this purpose a call for proposals [3] and a proposal package description [4] was issued by MPEG and the conditions for the test were defined [5]. The test included compression (coding efficiency) at bit rates of 2, 6, 16, 24, and 40 kbit/s, error resilience in the case of burst and random channel errors at bit rates of 6, 16, and 64 kbit/s, speed change at 2 and 6 kbit/s, and three different modes of scalability.

In response to this call various codecs were submitted. Due to the variety of functionalities, bit rates and types of audio material the proposed coding algorithms were based on different technologies ranging from specialized speech codecs for very low bit rates to general purpose audio codecs for high quality audio material. The candidate codecs evaluated in the subjective test are listed in Table 1, which also gives indications on key technologies and addressed functionalities [6].

The tests were carried out at different test laboratories in Japan, US, and Europe. The audio test signals were divided into different categories, the so called “item types” in order to allow more specific evaluation for music signals (item type “s”) and for speech signals which were subdivided into Japanese (item type “j”) and English/German (item type “e”). The two item types for speech signals were necessary for performing the listening tests as well in Japan as in Europe/US.

The selected test method for all bit rates below 64 kbit/s followed the ITU-T recommendation P.80, which specifies the so-called Absolute Category Ranking (ACR) [7], in which the Mean Opinion Score (MOS) for the quality is graded on a scale ranging from 1 (bad) to 5 (excellent). For bit rates of 64 kbit/s the test method according to ITU-R BS 1116 [8] was selected, in which the impairment of a distorted signal is graded with respect to the original signal. The exception were the signals decoded after introducing channel errors, which were assessed with the ACR method even for 64 kbit/s.

The statistical evaluation focused on finding significant factors which influenced the results. Of course the most interesting factor was the influence of the codecs, but one outcome was that some of the results from different test laboratories showed significant differences. Therefore the MPEG test report contains in addition to the total averages for all listeners the individual averages for all of the test laboratories [9]. However in this paper only total averages are presented in order to give an impression of the obtained quality and to allow a comparison with the included reference codecs. The reference codecs are listed in Table 2 which also shows the sampling rates of the audio material used at the different bit rates. The only exception were the input signals of the scalable codecs which had sampling frequencies corresponding to those of the highest bit rate. Since another significant factor is the item type, for the example of 16 kbit/s the item type specific averages are also presented here, while in the MPEG-4 test report they are included for all conditions.

Company / Institution	Basic Technologies	compression (kbit/s)	speed change (kbit/s)	error resilience (kbit/s)	scalability modes (kbit/s)
Alcatel / Philips / RAI	subband coding, ADPCM, adaptive Huffman coding	24, 40			64/6 64/24/6
AT&T	transform coding	24, 40			64/6
AT&T	waveform interpolation	2, 6			
Bosch / CSELT / MATRA (MAVT)	subband coding/LPC + enhancement for fine step scalability	6	6		64/24/6
INESC	subband/transform coding, Huffman coding, harmonic component extraction	16, 24, 40			
JVC	transform coding, VQ	40			
Matsushita	CELP + postprocessing for speed control + enhancement for error robustness	6	6	6	
Motorola	transform coding, Huffman coding + enhancement for error robustness	24, 40		64	64/24/6
NEC	transform coding + enhancement for scalability	24, 40			64/6 64/24/6
NEC	CELP with multi mode coding	6			64/6 64/24/6
NTT	transform coding, VQ + enhancement for error robustness	16, 24, 40		16, 64	64/24/6
NTT	CELP, pitch synchronous innovation + enhancement for error robustness	2, 6		6	16/6/2
NTT DoCoMo	CELP, pitch synchronous innovation + enhancement for error robustness			6	
Philips	CELP with efficient search strategies	16			
Samsung	subband coding, VQ				64/6
Sony	subband/transform coding + scalability with LPC				64/6
Sony IPC	subband/transform coding of LPC residual + enhancement for scalability				16/6/2
Sony IPC	LPC with harmonic vector excitation + enhancement for error robustness	2, 6	2,6	6	16/6/2
University of Erlangen	transform coding + scalability with low bit rate codecs	24, 40			64/6 64/24/6
University of Hannover / Deutsche Telekom	analysis/synthesis coding for individual spectral lines + enhancement for scalability	6, 16	6		

Table 1: Candidate codecs

target bit rate (kbit/s)	sampling frequency (kHz)	reference codec	operating bit rate of reference codec (kbit/s)
2	8	FS 1016	4.8
6	8	G729	8
16	16	G722	64
24	48	MPEG-2 Layer 3 (low samplig freq. mode)	24
40	48	MPEG-1 Layer 3	40
64	48	MPEG-1 Layer 3	64

Table 2: Reference codecs

During the evaluation of the test results for the VM development all proposals were clustered to three classes of codecs based on their key technologies:

- Class A: Codecs based on time/frequency mapping (i.e. transform/subband codecs)
- Class B: LPC-based analysis/synthesis codecs
- Class C: Codecs based on a parametric description of the audio signal

All of these three classes showed advantages for particular bit rates and/or functionalities as can be seen from the graphs in Figure 1 - Figure 5. The letters in the codec names in these graphs represent the classes. With the exception of the conditions with burst errors, random errors, and changed speed the mean values and 95% confidence intervals are shown. The so-called modulated noise reference units [10] which were included in the tests based on the ACR method are not shown here.

The tendencies in the behaviour of the best codecs of the different classes briefly can be summarized as follows: At a bit rate of 2 kbit/s codecs of class C achieved the best quality. At 6 kbit/s codecs of class B performed the best for speech, while for music there was no significant difference between the best codecs of class B and C. At 16 kbit/s again class B showed the best performance for speech (Figure 3b,c). For music however codecs of class A and C performed significantly better (Figure 3d). At bit rates of 24 kbit/s and above, only codec class A was represented. Furthermore other functionalities than compression have been investigated. Speed change and pitch change for example can be obtained easily for codecs of class C, while post-processing in the decoder is necessary for codecs of the other two classes. All these tendencies are summarized in Table 3 [11].

functionality	compression											speed		pitch
bit rate (kbit/s)	2			6			16			24	40	2	6	
item type	e	j	s	e	j	s	e	j	s					
codec class: A							o	o	+	+	+			
B	o	o		+	+	+	+	+	o				(+)	
C	+	+		o	o	+	-	-	+			+	+	+

Table 3: Tendencies in the performance of codec classes

In addition various implementations of bit rate/quality/bandwidth scalability have been proposed, also including combinations of codecs of different classes. Proposals addressing functionalities like low delay, low complexity, coding of multiple concurrent data streams, and pitch change were separately evaluated by an experts panel [12]. There also was a second call for proposals and a second submission phase. All proposals submitted in response to this call also were evaluated by an experts panel [13].

4 Structure of the Verification Model

Based on the results of the listening tests it was decided by MPEG that the audio VM shall support all three classes of codecs mentioned in Section 3 and that the VM provides the framework for combining the advantages of the different technologies [11]. Thus the MPEG-4 audio VM can be regarded as a library of tools, i.e. basic processing modules, and a framework which allows the construction of algorithms for given applications. From the evaluation of the test results a scenario was derived which indicates the interactions of the different coding schemes which correspond to the three classes (Figure 6). At bit rates between 2 and 8 kbit/s a parameter based coding scheme (class C) can be used. The bit rate of this scheme can be increased in small steps, which also lead to an improvement of the audio quality (small steps enhancement). For bit rates between 6 and 16 kbit/s an LPC based scheme (class B) is provided which also allows a small steps enhancement. At bit rates of 16 kbit/s and above a coding scheme which is based on time/frequency mapping (class A) can be selected. This module is adjustable in its output bit rate in larger steps and additionally allows the combination with the class B and C scheme for the implementation of scalability. A block diagram of the framework which implements this concept is shown in Figure 7. In addition to the codec cores for class A, B, and C and the enhancement modules a module for signal analysis and control, a pre-processing module, and a signal partitioning module is needed.

In the first steps of the VM implementation for each of the three classes a module based coding scheme description has been developed which allows the incorporation of all modules found to be promising during the evaluation of the test results. The interfaces between the individual modules have been defined in order to allow developers the preparation of alternative modules which can be compared to the currently selected implementations by core experiments [14]. A common framework has been implemented and a preliminary syntax for the control of the codec cores has been defined.

The block diagram of the class A core is shown in Figure 8. After a t/f-specific preprocessing the audio signal is decomposed into spectral components by the time/frequency mapping. The aim of the following feature extraction unit is to increase the efficiency of the quantization and coding unit for spectral components. The preprocessing, the feature extraction, and the quantization and coding units are controlled by a psychoacoustic model which is needed for the exploitation of properties of the human auditory system, e.g. by adjusting the quantizers to masking thresholds. The output bit stream is obtained by multiplexing the coded spectral components with the control information.

The main part of the class B core (Figure 9) is the LPC analysis/synthesis loop which consists of an excitation generator, an LPC synthesis filter, a spectral weighting filter for the synthesis error, and an error minimization unit. The audio input is fed through a preprocessing stage before it is given to the LPC analysis and quantization unit and to the analysis/synthesis loop. The coded excitation information and the quantized LPC parameters are combined to the output bit stream by a multiplexer.

Figure 10 shows the block diagram for the class C codec core which is based on a parametric signal representation from which decoder can re-synthesize an audio signal. The parameter extraction is applied to the audio signal after normalization by an LPC analysis stage. A pitch detection unit allows the extraction of parameters for harmonic structures. A spectral weighting unit derives control information for the quantization and coding of parameters. Control information and coded parameters are finally multiplexed in order to obtain the output bit stream.

The current status of the VM is defined by the specification of the modules from which the VM library is constituted [15]. Some of the key modules of the three classes are described in the following.

Since there is a requirement that MPEG-4 incorporates the MPEG-2 NBC codec for compression at bit rates of 64 kbit/s and above, the core for class A was designed in way that allows the implementation of an NBC compatible codec. Since NBC is based on time/frequency mapping there are two sets of modules for class A defined to be part of the VM library. The first set, implied by this requirement, contains for the time/frequency mapping an MDCT with window switching for long blocks (window length 2048 / frequency resolution 1024) and short blocks (256 / 128). The feature extraction contains backwards adaptive prediction and the quantization is implemented by scalar quantization with the extension for temporal noise shaping and by Huffman coding. The second set of modules contains a spectral normalization for the feature extraction and an interleaved vector quantization for the quantization and coding unit. A first check of the audio quality achieved by the application of the class A core based on these modules showed that it reaches a quality comparable to that of the best t/f based proposals in the test.

The biggest differences between the LPC based proposals can be found in the excitation generator. which therefore is one of the key components of the class B scheme. Currently it is based on a stochastic codebook for a bit rate of 6 kbit/s and on regular pulse excitation (RPE) for 16 kbit/s. The LPC analysis and quantization module includes functions like pitch estimation and conversions from LPC coefficients to line spectral frequencies (lsf). For the other modules there were no major differences between different proposals. A first check of a class B core implementation at 6 kbit/s showed that the achieved quality is only slightly lower than that of a G723.1 codec operating at 6.3 kbit/s. The RPE module is equivalent to that of the proposal which showed the best results for item type “e” in the 16 kbit/s test.

For class C the parameter extraction is the most important component, since it defines the type of parameters which are to be transmitted. Currently there are two implementations which are to be combined in the future. The first is based on a voiced/unvoiced decision and generates parameters for harmonic structures, i.e. high accuracy pitch and amplitudes of the corresponding harmonics, in the case of voiced signal segments and parameters for noise generation in the case of unvoiced signal segments. The second implementation extracts parameters for individual spectral lines, i.e. frequency, amplitude, and envelope. The class C core based on harmonic/noise parameters proved to give results equal to those of the best class C codec tested at 2 kbit/s, while the core based on individual lines behaves similar to a proposal which performed quite well for the music items in the 6 and 16 kbit/s test.

5 Possible Core Experiments

This Section gives some indications on alternative components of various proposals which are not included in the current VM library, but which might improve the coding efficiency or add functionality or flexibility.

- Framework
    - partitioning: signal classification
- Class A
    - time/frequency mapping (t2f): t2f with multiple overlap and selectable frequency resolution, complex t2f, cascaded t2f, low delay t2f
    - feature extraction: tonal component extraction, backwards adaptive block prediction
    - quantization and coding: vector quantization, small step scalable vector quantization, alternative quantization and entropy coding
    - other: (hardware) scalability
- Class B
    - different excitation modes at different bit rates, e.g. 6, 10, 16 kbit/s
- Class C
    - decomposition into slowly and rapidly evolving waveforms

After the definition of profiles for error resilience core experiments for error protection schemes should be carried out.

6 Conclusions

In this paper the development and the current status of the MPEG-4 audio verification model are described. The results of the listening tests carried out by MPEG show that based on the proposed coding techniques improvements with respect to existing standards seem very likely. These improvements can consist of increased coding efficiency or added functionalities. The basic structure of the VM is shown as well as the more detailed block diagrams for the three supported codec classes which are: codecs based on time/frequency mapping, LPC based analysis/synthesis codecs, and codecs based on a parametric signal representation. The current implementations of corresponding processing modules are also described. With the current version of the VM the coding efficiency under the most conditions is comparable to that of the best proposals submitted to the first subjective test. Further improvements are expected from the collaborative phase which allows optimization based on core experiments. However several extensions to the current version of the VM are still necessary. For example some of the enhancement modules are not yet implemented which allow a finer adjustment of the bit rate and which are needed for certain modes of scalability. Error correction and concealment are also not yet addressed and speed/pitch change is only partly implemented. Future work also will include multichannel coding, coding of multiple concurrent data streams, low delay modes, and low complexity modes. Another major issue which needs to be addressed is to match the syntax definition with the requirements for the MPEG-4 Syntax Description Language (MSDL) [16].

7 References

[1] ISO/IEC JTC1/SC29/WG11/N1191, “MPEG-4 Development Procedures”, Florence, March 1996.

[2] ISO/IEC JTC1/SC29/WG11/N1307, “Text of CD 13818-7 Audio NBC”, Tampere, July 1996.

[3] ISO/IEC JTC1/SC29/WG11/N997, “MPEG-4 Call for Proposals”, Tokyo, July 1995.

[4] ISO/IEC JTC1/SC29/WG11/N998, “MPEG-4 Porposal Package Description”, Tokyo, July 1995.

[5] ISO/IEC JTC1/SC29/WG11/N999, “MPEG-4 Testing and Evaluation Procedures Document”, Tokyo, July 1995.

[6] Contin, L., Edler, B., Meares, D., Schreiner, P., “Tests on MPEG-4 Audio Codec Proposals”, Signal Processing: Image Communication Journal, October 1996.

[7] ITU-T, “Methods for subjective determination of transmission quality”, ITU-T Recommendation P80, 1993.

[8] ITU-R, “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems”, ITU-R Recommendation BS-1116, November, 1993.

[9] ISO/IEC JTC1/SC29/WG11/N1144, “ MPEG-4 Audio Test Results (MOS Tests)”, Munich, January 1996.

[10] ITU-T, “Modulated Noise Reference Unit (MNRU)”, ITU-T Recommendation P81,1993.

[11] ISO/IEC JTC1/SC29/WG11/N1143, “Preliminary Draft of MPEG-4 Audio VM”, Munich, January 1996.

[12] Brandenburg, K., “Report on Activities of the Ad-hoc Group on the Evaluation of Tools for Non-tested Functionalities of Audio Submissions to MPEG-4”, ISO/IEC JTC1/SC29/WG11/M0473, Dallas, October 1995.

[13] Pan, D., “Report of the Adhoc Group on the Evaluation of New Audio Submissions to MPEG-4”, ISO/IEC JTC1/SC29/WG11/M680, Munich, January 1996.

[14] ISO/IEC JTC1/SC29/WG11/N1214, “Draft MPEG-4 Audio Verification Model”, Florence, March 1996.

[15] ISO/IEC JTC1/SC29/WG11/N1345, “MPEG-4 Audio Verification Model 1.1”, Tampere, July 1996.

[16] ISO/IEC JTC1/SC29/WG11/N1381, “MSDL Specification. Version 1.2”, Tampere, July 1996.

Figures

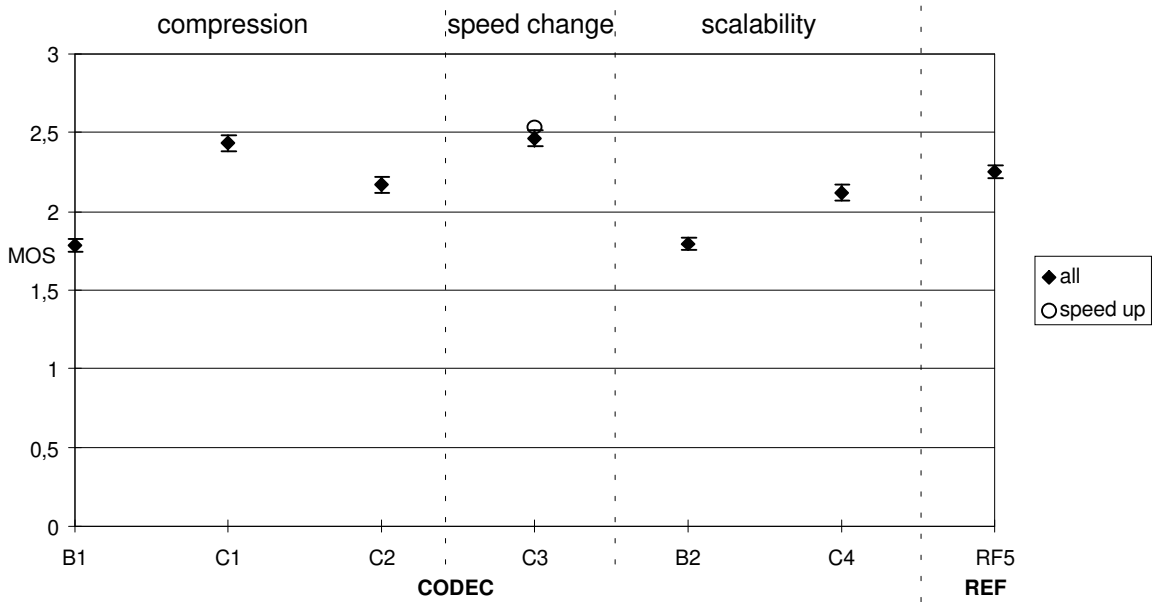

Figure 1: Results of the 2 kbit/s tests (average over all items).

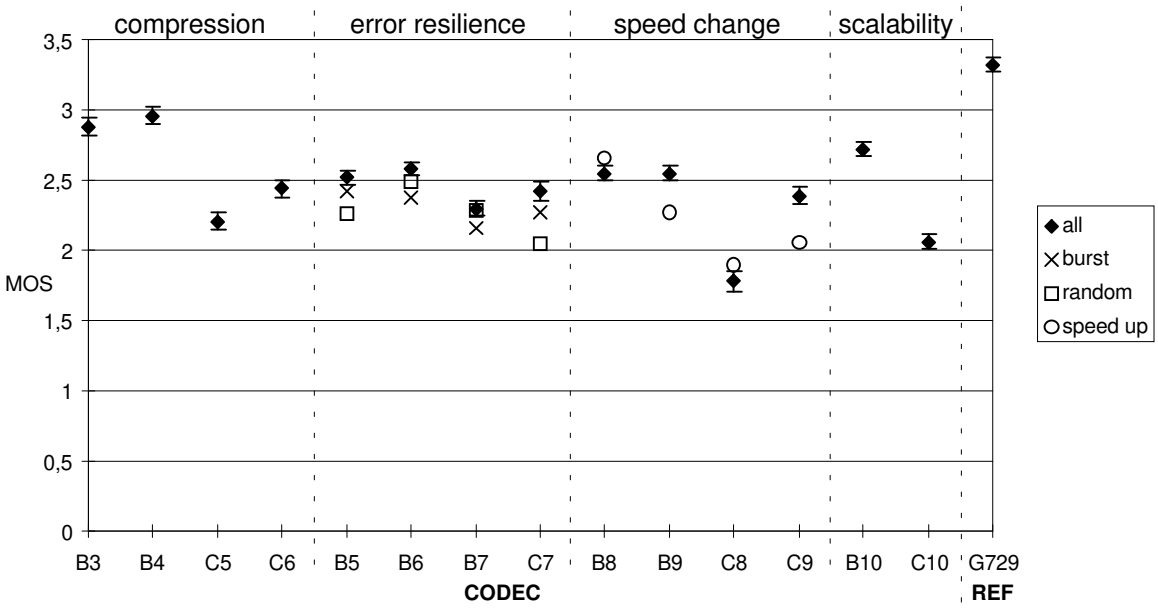

Figure 2: Results of the 6 kbit/s tests (average over all items).

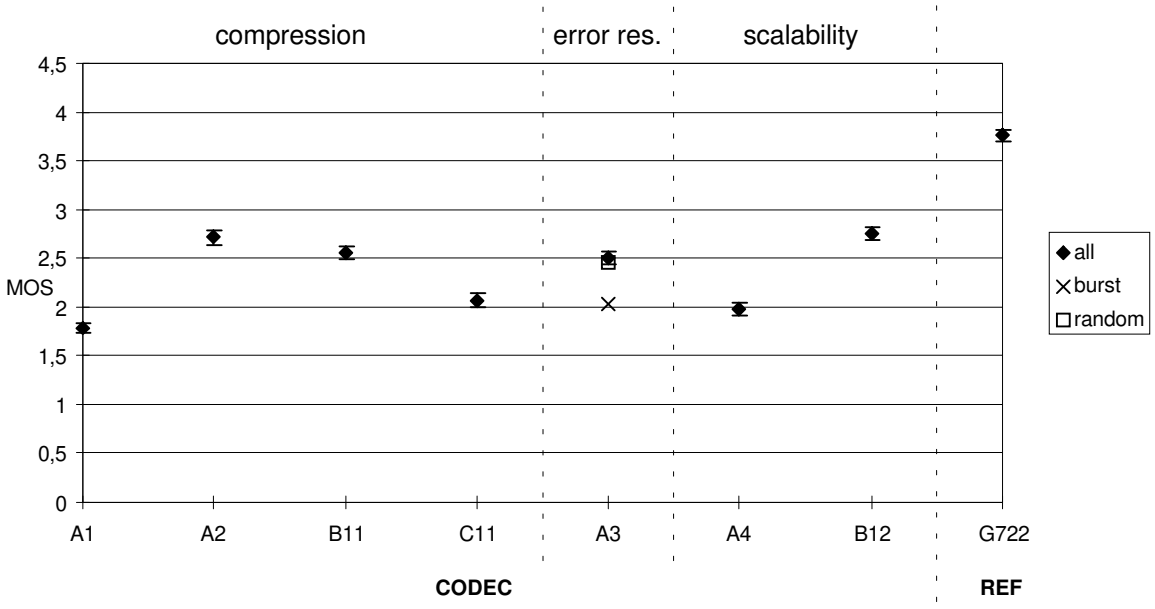

Figure 3a: Results of the 16 kbit/s tests (average over all items).

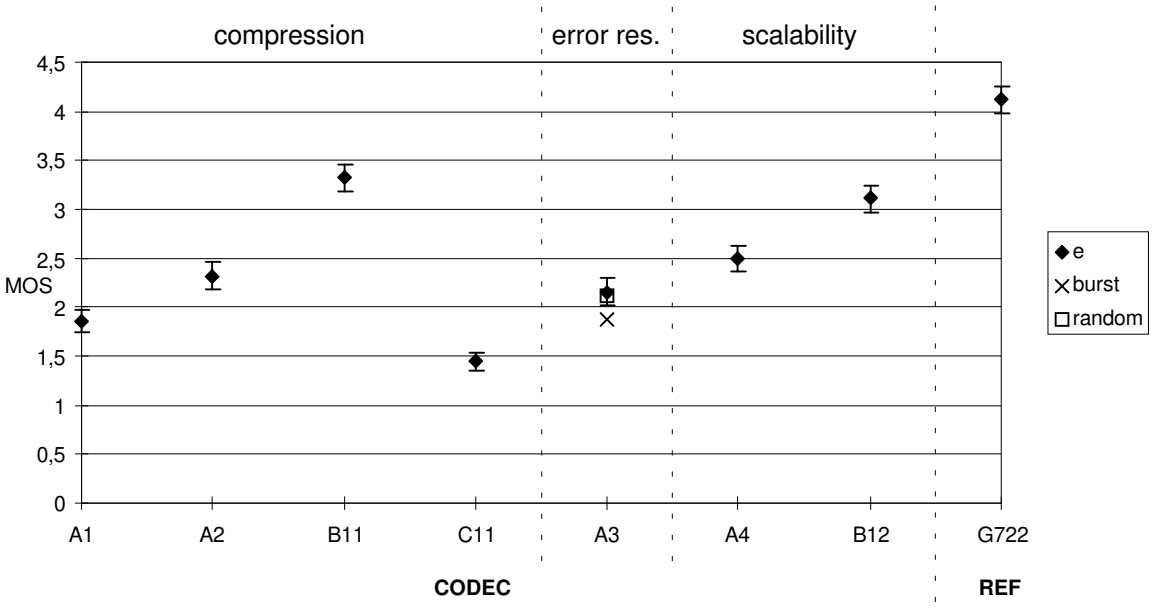

Figure 3b: Results of the 16 kbit/s tests (average over all “e” items).

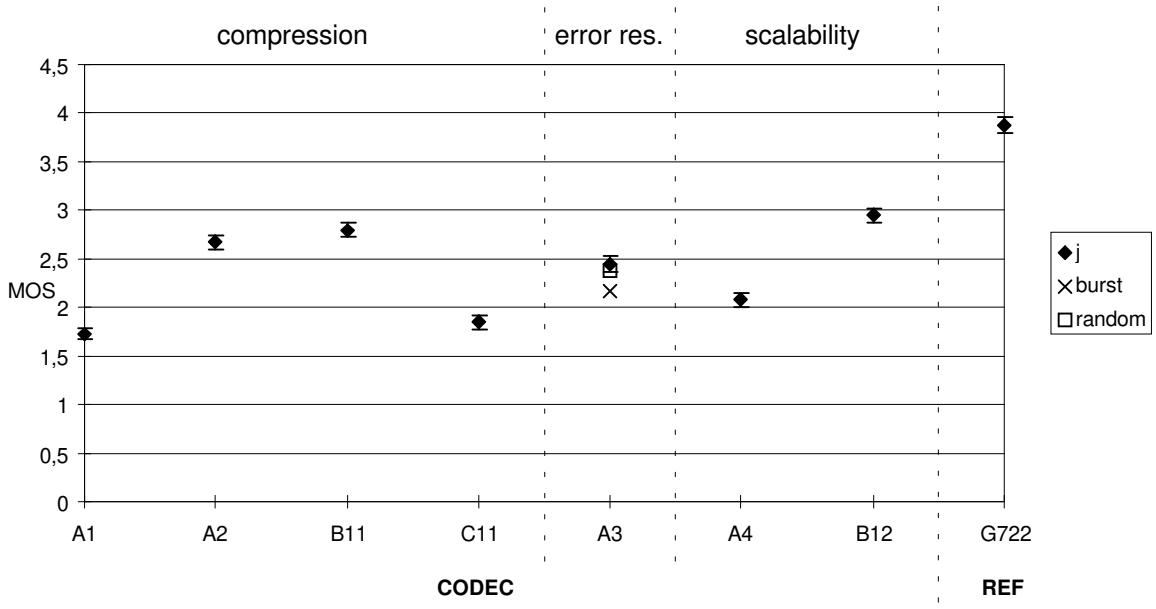

Figure 3c: Results of the 16 kbit/s tests (average over all “j” items).

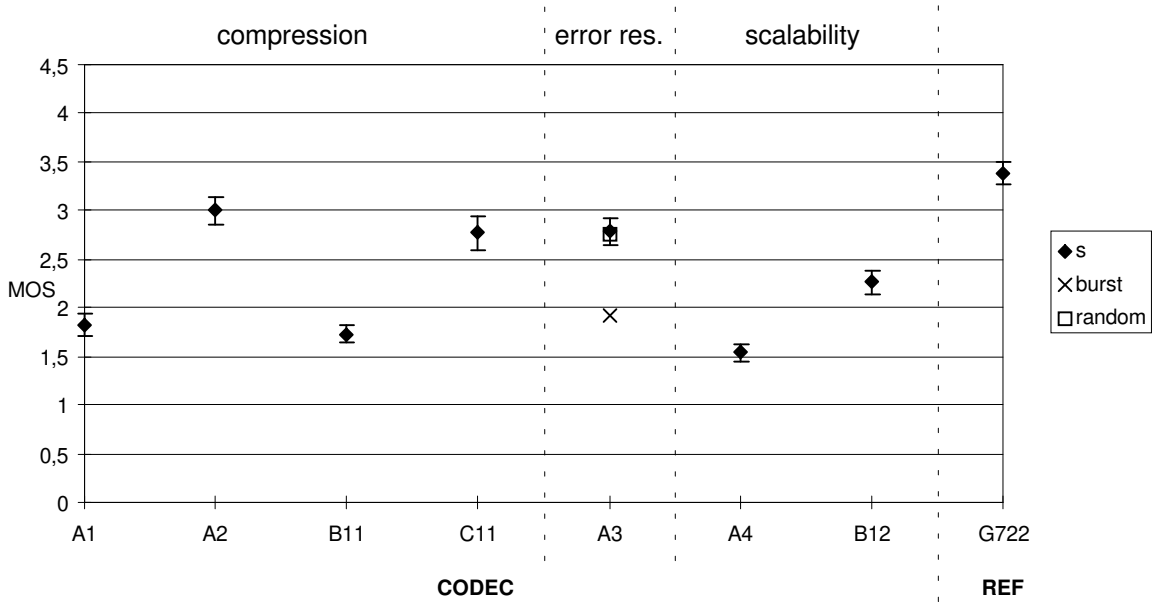

Figure 3d: Results of the 16 kbit/s tests (average over all “s” items).

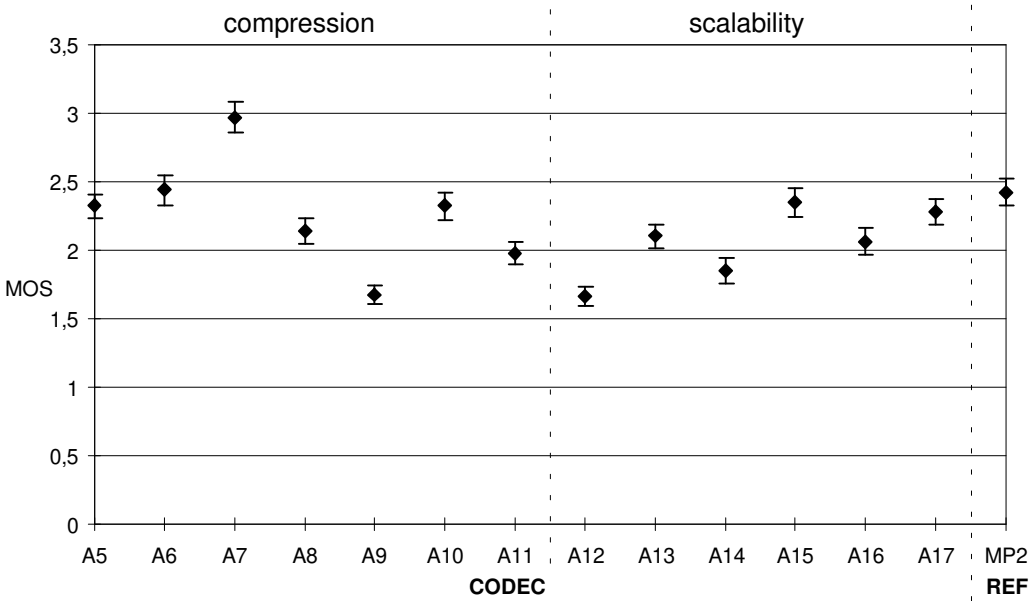

Figure 4: Results of the 24 kbit/s tests (average over all items).

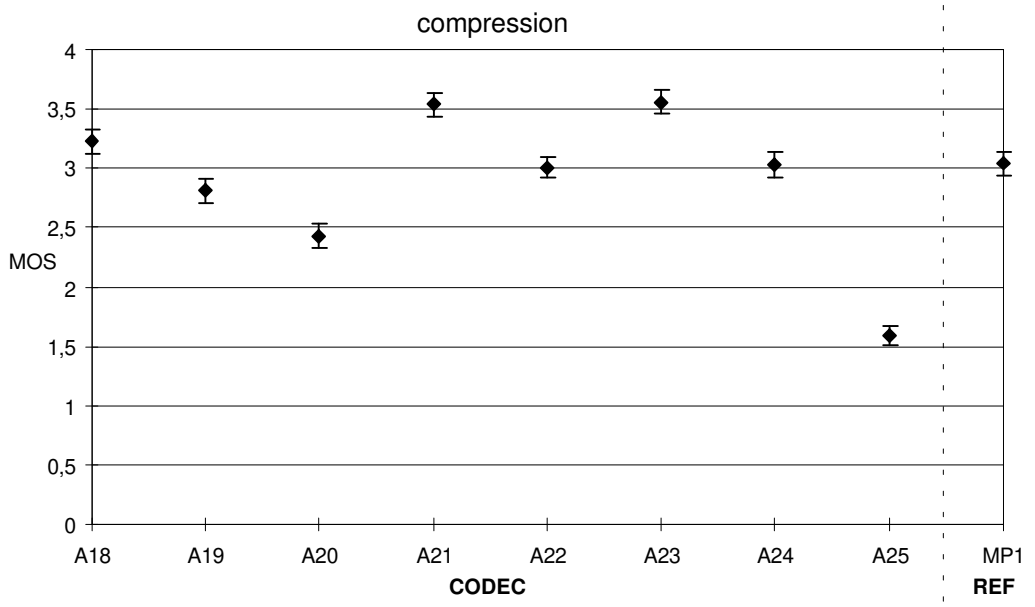

Figure 5: Results of the 40 kbit/s tests (average over all items).

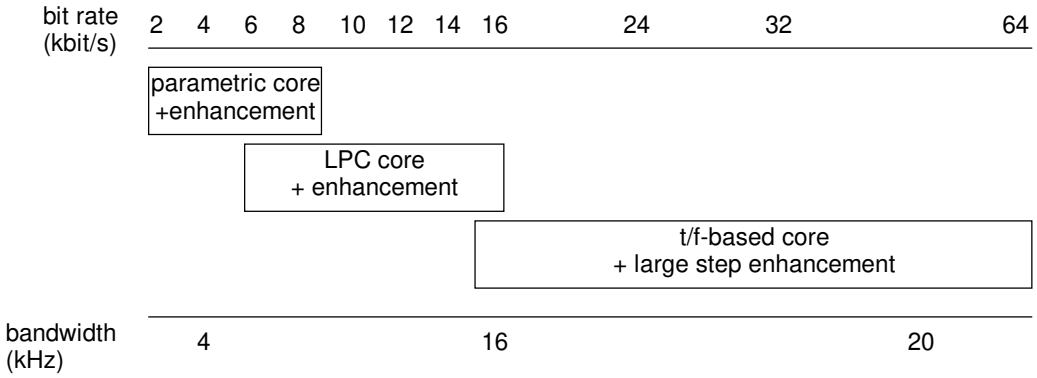

Figure 6: Scenario for the application of different codec classes.

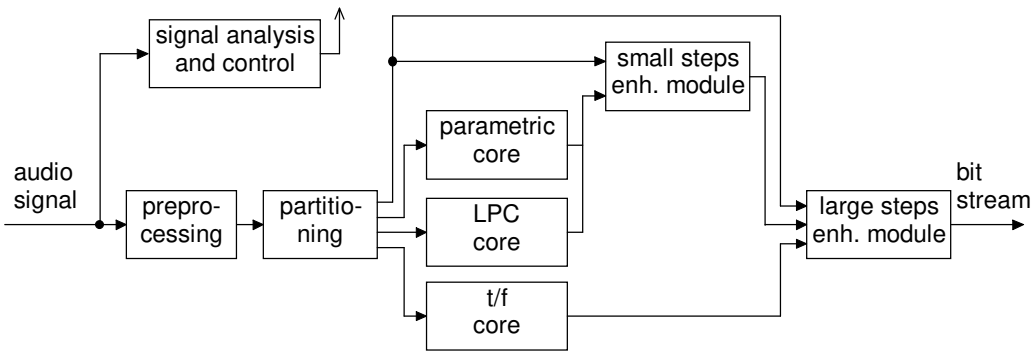

Figure 7: Block diagram of the MPEG-4 Audio VM framework.

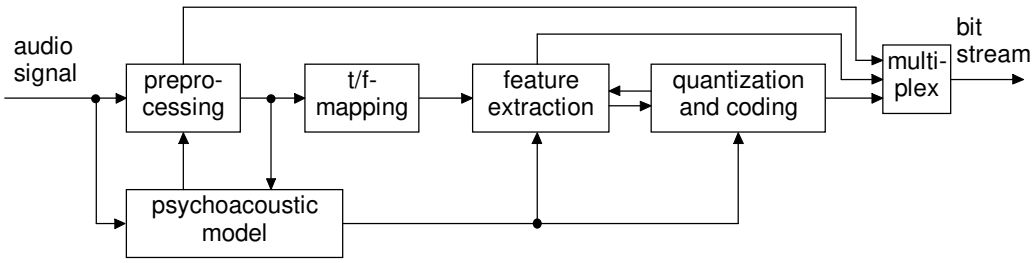

Figure 8: Block diagram for class A (t/f-based) core.

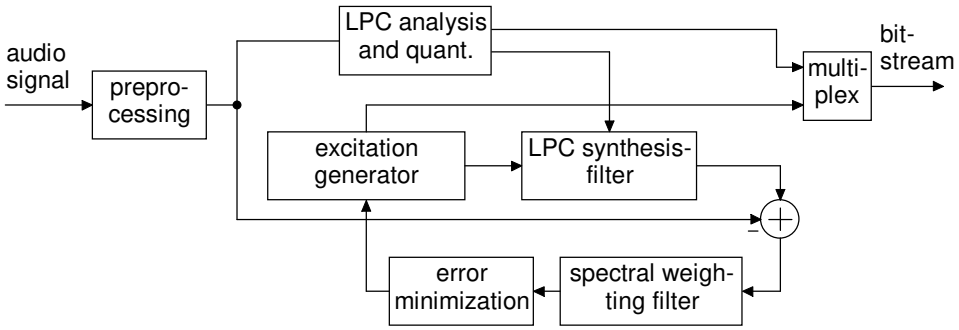

Figure 9: Block diagram for class B (LPC-based) core.

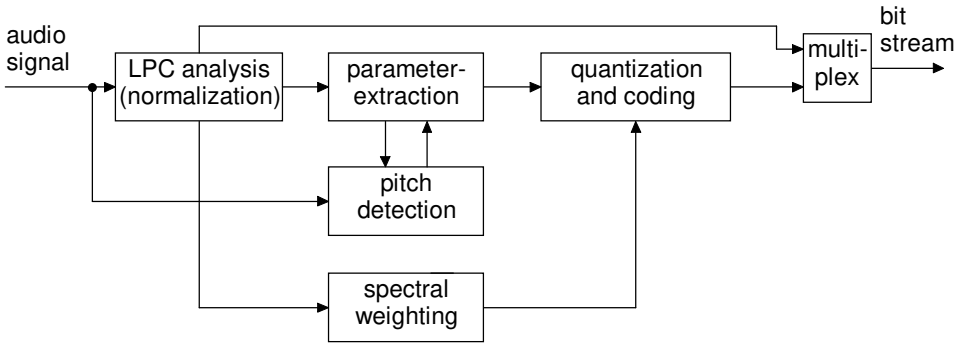

Figure 10: Block diagram for class C (parameter based) core.