INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC 1/SC 29/WG 11 N3075
December 1999

Source:	**Audio Subgroup**
Title:	**Report on the MPEG-4 Audio Version 2 Verification Test**
Status:	**Approved**
Authors:	**Ralph Sperschneider (FhG), Frank Feige (T-Nova), Schuyler Quackenbush (AT&T)**

Summary

The MPEG-4 Audio Version 2 coding tools have undergone a performance verification test for coding of monophonic audio signals in the range of 6 kbit/s to 64 kbit/s and stereophonic audio signals in the range of 64 kbit/s to 96 kbit/s. The coding tools tested were Harmonic and Individual Lines plus Noise (HILN) coding, Bit Sliced Arithmetic Coding (BSAC), Low Delay Advanced Audio Coding (AAC LD) and the Error Robustness tools comprising Error Resilience (ER), and Error Protection (EP). It was found that, relative to Version 1 tools, Version 2 tools provide new capabilities while still providing comparable audio quality and comparable levels of compression. New capabilities evaluated as part of these tests are parametric signal representation (allowing independent speed and pitch modification), fine step bit rate scalability, very low communications delay, and robustness to channel errors.

1 Table of Contents

2 Introduction

MPEG-4 Version 2 is the name given to technology in Amendment 1 of MPEG-4 (ISO/IEC 14496). Although it is an amendment, Version 2 is more correctly viewed as technology that required more time to develop and hence was not available at time that ISO/IEC 14496 was issued as an international standard. The purpose of the tests reported on here is to verify that Version 2 tools bring valuable technology to the MPEG-4 standard. The figure of merit in the test is subjective audio quality. This, plus each tool's features and capabilities, permit system developers to better judge the merit of the technology as a basis for future applications.

The technology tested was Harmonic and Individual Lines plus Noise (HILN) coding, Bit Sliced Arithmetic Coding (BSAC), Low Delay Advanced Audio Coding (AAC LD) and the Error Robustness tools comprising Error Resilience (ER) and Error Protection (EP). While the Version 2 technology provides compression, it is most often compression in conjunction with other valuable features, such as very low bit rate (for HILN), very low delay (for AAC LD), fine step bit rate scalability (for BSAC) or robustness to bit stream errors (for ER and EP tools). The ER and EP tools are valuable in systems in which compressed audio information must be transmitted over error-prone channels. These may be radio channels that incur bit or byte errors, or packet channels that incur lost (or late) packets. The increasing importance of wireless communications and the Internet make these tools particularly valuable.

In this document the names of the following Audio object types are used to identify the different codecs (for details see [n3058]):

object type ID	Audio object type	version	description
1	AAC main	1	Advanced Audio Coding in main configuration
3	AAC SSR	1	Advanced Audio Coding in scalable sampling rate configuration
8	CELP	1	Code Excited Linear Prediction
12	TTSI	1	Text to speech interface
7	TwinVQ	1	Transform Domain weighted interleave Vector Quantization
17	ER AAC LC	2	Error Resilient Advanced Audio Coding with Low Complexity
23	ER AAC LD	2	Error Resilient Advanced Audio Coding with Low Delay
20	ER AAC scalable	2	Error Resilient scalable Advanced Audio Coding
22	ER BASC	2	Error Resilient Bit Sliced Arithmetic Coding
26	ER HILN	2	Error Resilient Harmonic and Individual Lines plus Noise
25	ER HVXC	2	Error Resilient Harmonic Vector Excitation Coding
21	ER TwinVQ	2	Error Resilient Transform Domain weighted interleave Vector Quantization

Table 2-1: Audio object types considered within this test report

The set of new tools provided by MPEG-4 Audio Version 2 is listed below:

New codecs:
- ER HILN, Parametric (ER HVXC + ER HILN)
- ER AAC LD
- ER BSAC

Codec extensions:
- Silence compression for ER CELP
- Variable rate coding for ER HVXC at 4 kbit/s

Error robustness:

- EP tool
- Error resilient bit stream syntax for all Version 1 object types (except of AAC main, AAC SSR, TSSI and structured audio related object types)
- Error resilience tools for ER AAC LC, ER AAC LTP, ER AAC scalable, and ER AAC LD
- Error resilience mode for ER BSAC

Out of this pool, the following Version 2 object types have been evaluated in this test:

- ER HILN (Session A1)
- ER BSAC (Session A2)
- ER AAC LD (Session A3)
- Error robustness applied to ER AAC LC and ER TwinVQ (Session A4)

No per-item tuning was permitted on any of the codecs involved in these verification tests.

3 Codecs under Test

During the Vancouver MPEG meeting it was decided to test the following Version 2 coding tools in three distinct sessions: ER HILN, ER BSAC and ER AAC LD. It was also decided to test in a separate session ER and EP tools as they apply to ER AAC LC and ER TwinVQ. The four sessions are designated A1, A2, A3, and A4.

The tables in this chapter indicate the parameters for the respective codec under test, the test method, and the reference codec. The reference codec serves as an anchor in the test, permitting results from this test to be more easily compared to that of previous tests in which the same reference codec was also tested.

3.1 Session A1 – ER HILN

Codec under test	Reference Codec	Test method
ER HILN 6 kbit/s @ 16 kHz (mono)	TwinVQ 6 kbit/s @ 16 kHz (mono)	BS.1284 quality scale, R/A R: band limited to 8 kHz
ER HILN scalable 6 kbit/s @ 16 kHz (mono) based on scalable configuration: 6 kbit/s @ 16 kHz (mono) + 10 kbit/s @ 16 kHz (mono)	TwinVQ 6 kbit/s @ 16 kHz (mono)	BS.1284 quality scale, R/A R: band limited to 8 kHz
ER HILN 16 kbit/s @ 16 kHz (mono)	AAC main 16 kbit/s @ 22.05 kHz (mono)	BS.1284 quality scale, R/A R: band limited to 8 kHz
ER HILN scalable 16 kbit/s @ 16 kHz (mono) based on scalable configuration: 6 kbit/s @ 16 kHz (mono) + 10 kbit/s @ 16 kHz (mono)	AAC main 16 kbit/s @ 22.05 kHz (mono)	BS.1284 quality scale, R/A R: band limited to 8 kHz

Table 3-1: Overview of session A1 - ER HILN

No.	Codec	Sampling rate
1	ER HILN 6 kbit/s	16 kHz
2	ER HILN 16 kbit/s	16 kHz
3	ER HILN (6 +10) kbit/s	16 kHz
4	TwinVQ 6 kbit/s	16 kHz
5	AAC main 16 kbit/s	22.05 kHz

Table 3-2: Codecs for session A1 (mono)

ER HILN Codec Setup

The following information is compiled from [m5045].

Three different bit streams were prepared for each of the items:

- 6 kbit/s single layer ER HILN bit stream
- 16 kbit/s single layer ER HILN bit stream
- 6 kbit/s base layer + 10 kbit/s extension layer scalable ER HILN bit stream

The following encoder configuration was used:

sampling rate:	16 kHz
bandwidth:	8 kHz
number of channels:	1 (mono)
frame size:	32 ms
bit reservoir size:	384 bits (= 64 ms) for 6 kbit/s single layer bit streams 1024 bits (= 64 ms) for 16 kbit/s single layer bit streams no bit reservoir for 6 kbit/s +10 kbit/s scalable bit streams

TwinVQ Codec Setup

sampling rate	16 kHz
number of channels	1
bandwidth	2.8 kHz
bit rate	6 kbit/s
frame size	1024 points

Table 3-3: Parameters on Source coding for TwinVQ reference codec

AAC main Codec Setup
The following information is compiled from [m4998].

For comparison with the 16 kbit/s ER HILN, the reference AAC encoder operated at 16 kbit/s @ 22.05 kHz (encoder-internal resampling from supplied 16 kHz wave files). It was configured to produce bit stream payloads of AAC main object type.

3.2 Session A2 – ER BSAC

Codec under test	Reference Codec	Test method
ER BSAC 96 kbit/s @ 32 kHz (stereo)	AAC main 96 kbit/s @ 32 kHz (stereo)	BS.1284 Quality scale, R/A/R/A
ER BSAC 88 kbit/s @ 32 kHz (stereo) derived from configuration 96 kbit/s @ 32 kHz (stereo)	AAC main 96 kbit/s @ 32 kHz (stereo)	BS.1284 Quality scale, R/A/R/A
ER BSAC 80 kbit/s @ 32 kHz (stereo) derived from configuration 96 kbit/s @ 32 kHz (stereo)	AAC main 96 kbit/s @ 32 kHz (stereo)	BS.1284 Quality scale, R/A/R/A
ER BSAC 72 kbit/s @ 32 kHz (stereo) derived from configuration 96 kbit/s @ 32 kHz (stereo)	AAC main 96 kbit/s @ 32 kHz (stereo)	BS.1284 Quality scale, R/A/R/A
ER BSAC 64 kbit/s @ 32 kHz (stereo) derived from configuration 96 kbit/s @ 32 kHz (stereo)	AAC main 64 kbit/s @ 32 kHz (stereo)	BS.1284 Quality scale, R/A/R/A

Table 3-4: Overview of Session A2 - ER BSAC

No.	Codec	Sampling rate
1	ER BSAC 64 kbit/s	32 kHz
2	ER BSAC 72 kbit/s	32 kHz
3	ER BSAC 80 kbit/s	32 kHz
4	ER BSAC 88 kbit/s	32 kHz
5	ER BSAC 96 kbit/s	32 kHz
6	AAC main 64 kbit/s	32 kHz
7	AAC main 96 kbit/s	32 kHz

Table 3-5: Codecs for session A2 (stereo)

ER BSAC Codec Setup
The ER BSAC encoder operated at 96 kbit/s @ 32 kHz (resampling from supplied 48 kHz stereo wave files using ResampAudio from the AFsp library).

Following bit streams are derived from the 96 kbit/s bit stream.

88 kbit/s @ 32 kHz
80 kbit/s @ 32 kHz
72 kbit/s @ 32 kHz
64 kbit/s @ 32 kHz

AAC main Codec Setup
The following information is compiled from [m4998].

For comparison with the 96 kbit/s ER BSAC, the reference AAC encoder operated at 96 kbit/s @ 32 kHz (encoder-internal resampling from supplied 48 kHz stereo wave files). It was configured to produce bit stream payloads of AAC main object type.

For comparison with the 64 kbit/s ER BSAC, the reference AAC encoder operated at 64 kbit/s @ 32 kHz (encoder-internal resampling from supplied 48 kHz stereo wave files). It was configured to produce bit stream payloads of AAC main object type.

3.3 Session A3 – ER AAC LD

Codec under test	Reference Codec	Test method
ER AAC LD 64 kbit/s @ 48 kHz (mono) 20 ms delay	AAC main 56 kbit/s @ 44.1 kHz (mono)	BS.1284 quality scale, R/A/R/A R: full band original
ER AAC LD 32 kbit/s @ 32 kHz (mono) 30 ms delay	AAC main 24 kbit/s @ 24 kHz (mono) G.722 64 kbit/s @ 16 kHz (mono) CELP 24 kbit/s @ 16 kHz (mono)	BS.1284 quality scale, R/A/R/A R: band limited to 8 kHz

Table 3-6: Overview of Session A3 @ ER AAC LD

No.	Codec	Sampling rate
1	ER AAC LD 64 kbit/s	48 kHz
2	AAC main 56 kbit/s	44.1 kHz

Table 3-7: Codecs for session A3 – 64 kbit/s (mono)

No.	Codec	Sampling rate
1	ER AAC LD 32 kbit/s	32 kHz
2	AAC main 24 kbit/s	24 kHz
3	CELP 23.8 kbit/s	16 kHz
4	ITU-T G.722 64 kbit/s	16 kHz

Table 3-8: Codecs for session A3 – 32 kbit/s (mono)

ER AAC LD Codec Setup
The following information is compiled from [m4998].

At a bit rate of 64 kbit/s, the ER AAC LD encoder operated at an internal sampling rate of 48 kHz, 480 lines of spectral resolution and no use of the bit reservoir. This corresponds to an overall algorithmic delay of 20 ms.

At a bit rate of 32 kbit/s, the ER AAC LD encoder operated at an internal sampling rate of 32 kHz, 480 lines of spectral resolution and no use of the bit reservoir. This corresponds to an overall algorithmic delay of 30 ms.

Because error robustness capabilities are not subject of this test, both encoders used the raw data stream syntax as defined for AAC in Version 1 instead of the error resilient syntax.

AAC main Codec Setup
The following information is compiled from [m4998].

For comparison with the 64 kbit/s ER AAC LD, the reference AAC encoder operated at 56 kbit/s @ 44.1 kHz (encoder-internal resampling from supplied 48 kHz wave files). It was configured to produce bit stream payloads of AAC main object type.

For comparison with the 32 kbit/s ER AAC LD, the reference AAC encoder operated at 24 kbit/s @ 24 kHz (encoder-internal resampling from supplied 32 kHz wave files). It was configured to produce bit stream payloads of AAC main object type.

G.722 Codec Setup
16 kHz sampling rate versions of the PCM test samples were used as input to the ITU-T G.722 wideband speech coder found in the STL provided by ITU-T. 64 kbit/s bit rate was chosen, and this was the only adjustable parameter of the coder. The bit streams were decoded using the decoder part of the same G.722 codec. The output of the decoder was again PCM with 16 kHz sampling rate.

CELP Codec Setup

bit rate	23.8 kbit/s
sampling rate	16 kHz
frame length	10 ms
algorithmic delay	15 ms (including 5 ms look ahead)
excitation mode	MPE
scalability	no bit rate scalability, no bandwidth scalability
fine rate control	none

Table 3-9: Parameters on Source coding for CELP reference codec

3.4 Session A4 – Error Robustness

Codec under test	Reference Codec	Test method
ER AAC LC (incl. ER tools) 96 kbit/s @ 32 kHz (stereo) EP Tool critical error condition	ER AAC LC (incl. ER tools) 96 kbit/s @ 32 kHz (stereo)	MUSHRA (see section 5.2)
ER AAC LC (incl. ER tools) 96 kbit/s @ 32 kHz (stereo) EP Tool very critical error condition	ER AAC LC (incl. ER tools) 96 kbit/s @ 32 kHz (stereo)	MUSHRA (see section 5.2)
ER TwinVQ 16 kbit/s @ 32 kHz (mono) EP Tool critical error condition	ER TwinVQ 16 kbit/s @ 32 kHz (mono)	MUSHRA (see section 5.2)
ER TwinVQ 16 kbit/s @ 32 kHz (mono) EP Tool very critical error condition	ER TwinVQ 16 kbit/s @ 32 kHz (mono)	MUSHRA (see section 5.2)

Table 3-10: Overview of Session A4 - Error Robustness

ER AAC LC Codec Setup
The following information is compiled from [m4998].

At a bit rate of 96 kbit/s, the AAC encoder operated at an internal sampling rate of 32 kHz (encoder-internal resampling from supplied 48 kHz wave files). It was configured to produce bit stream payloads of ER AAC LC object type.

A Version 1 to Version 2 AAC transcoder was used to translate bit stream payloads of AAC LC object type to those of ER AAC LC object type. It was configured to apply noiseless AAC error resilience tools (HCR and VCB11).

The EP tool was used to produce unequal error protected bit stream payloads. Its configuration was as follows:

- Rearrange error sensitivity category instances as follows: 0a, 1a, 1b, 2a, 2b, 3a, 3b, 4a, 4b.
- Discard empty instances (done by using several predefinition sets)
- Joint application of FEC using RS(255-l, 245-l) on instances 0a, 1a, 1b
- Intra-class interleaving for instances 4a, 4b and FEC protected part
- EP header interleaving
- Bit stuffing (byte alignment)

The total overhead added for error robustness is 9.5 % (2 % for ER & 7.5 % for EP).

The following concealment procedures are used on decoder site:

- If the current frame is lost or side information CRC is erroneous, the whole MDCT spectrum is concealed for the appropriate channel.
- Particular MDCT lines are concealed if they are detected to be erroneous by one of the ER tools.

A combination of noise substitution and prediction in conjunction with energy interpolation is used as concealment technique. The selection of the appropriate concealment method depends on the signal characteristics. A delay of one frame is inserted due to the concealment. If a multiple frame loss occurs the reconstructed spectra are attenuated.

ER TwinVQ Codec Setup
The following information is compiled from [m5051].

Concatenated input material	
	Source encoding
Flexmux bit stream 66 Byte/frame	↓
	Header removing
Raw bit stream 64 Byte/frame	↓
	EP tool encoding
Protected bit stream + side information 75 Byte/frame	↓
	Error Insertion/Mux-demux
Distorted bit stream + frame erasure/CRC information	↓
	EP tool decoding
Reconstructed bit stream 64 Byte/frame	↓
	Header merging
Flexmux bit stream 66 Byte/frame	↓
	Source decoding + concealment
Output signal	

Table 3-11: Signal generation process

sampling rate	32 kHz
number of channels	1
bit rate (source)	16 kbit/s
bit rate (redundancy)	2.75 kbit/s (17.2 %)
frame size	1024 samples, 32 ms
bit/frame	512
Number of UEP classes	3
Byte/frame	64

Table 3-12: Parameters for source coding

Number of configurations	1	
Interleave mode	1	YES for class 1 and 2
Bit stuffing	0	
Number of classes	3	
	0 0 0	no escape
Number of source bits for class 1	12	Flags and MSB of gain
Redundancy rate for class 1	16	Rate 24/8
CRC bits for class 1	4	
	0 0 0	no escape
Number of source bits for class 2	22	Parameters
Redundancy rate for class 2	11	Rate 19/8
CRC bits for class 2	9	
	0 0 0	no escape
Number of source bits for class 3	478	Index for MDCT VQ
Redundancy rate for class 3	0	No protection
CRC bits for class3	0	No CRC
	0	no header

Table 3-13: UEP configuration

The following concealment procedures are used on decoder site:

- If the current frame is lost or the first class CRC is erroneous, waveform is extrapolated from the previous frame in the time domain.
- If only the second class CRC is erroneous, reconstructed spectrum is attenuated. Especially, when the frame energy has significantly increased, spectrum gain is reduced so that the frame energy is smaller than that of the previous frame.
- If the previous frame is lost or erroneous, frame gain is slightly attenuated even though the current frame has no errors.
- No additional delay is introduced due to error concealment.

Error insertion and multiplexing / demultiplexing are applied to the error protected bit streams. Based on the frame erasure information and the CRC information, concealment processes were carried out, and there was no additional delay due to the concealment process.

Channel Setup

Transmission simulation is done on a continuous sequence. Due to this all (eight) items are concatenated prior to encoding. The error robust encoded data is processed by a multiplex layer to produce a bit stream ready for error insertion. The error pattern is applied to this bit stream. After decoding the sequence is split again into the eight items, which are then graded separately.

For multiplexing, error insertion and de-multiplexing the wireless system model as shown below is used:

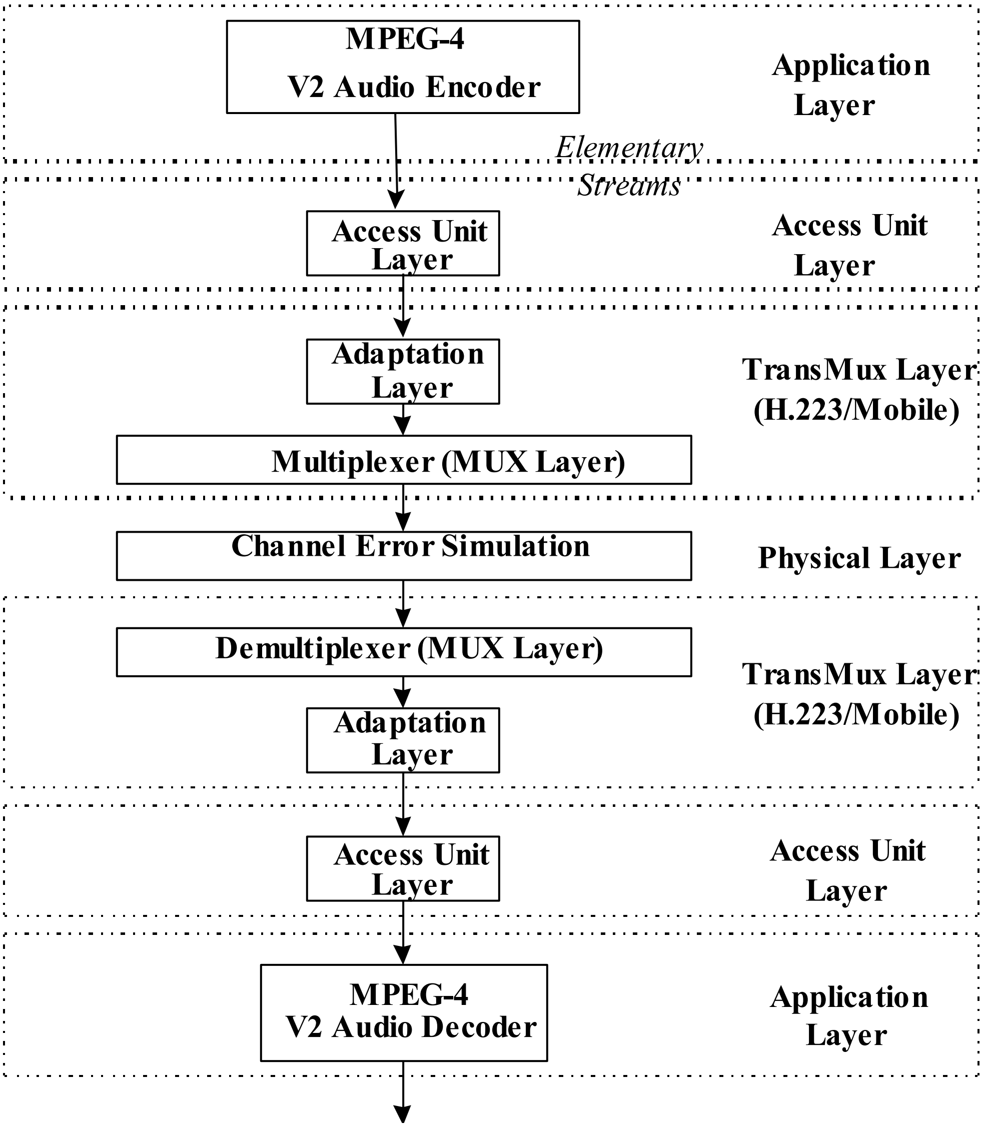


Figure 3-1: Wireless System Model used for Error Robustness Tests

Application Layer	MPEG-4 Audio Encoder/Decoder
Access Unit Layer	Sync Layer with “Null SL-packet header”
TransMux Layer	H.223/Mobile : a mobile extension of ITU’s standard for videophone multiplexer
Physical Layer	10 ms burst error: typical mobile channel condition 1 ms burst error: critical mobile channel condition

Table 3-14: Layers used in Error Robustness Test

As the simplest model of the Sync Layer, the following assumption is employed:

- One Access Unit corresponds to one Audio Packet.
- One Access Unit is mapped into one AL-PDU
- No SL-packet header is used, assuming ‘Null SL-packet Header’ with a configuration of “predefine = 0x01”

H.223 mobile mode 2 (H.223 with its Annex B, see [h223B]) was selected as a TransMux, amongst a variety of H.223 and its extensions. Here, the header information in multiplexed packets is strongly protected, but its payload, i.e. audio packets is not protected at all.

Major parameters used for the TransMux are as follows:

- **Adaptation Layer for Audio**
  Amongst three adaptation layers defined in H.223 main body, AL2 (type 2) is used in this verification test. The unit of the transmission exchanged with the codec is called “AL-SDU”. The encoder determines the size of AL-PDU, and this layer guarantees the boundary of this AL-SDU. The AL-SDU is aligned with AL-PDU of Access Unit Layer, i. e. it is aligned with audio packet.
- **Control**
  Unlike audio, the control information was transmitted only during the initialisation phase, and thus the transmission of this information is not necessary for the realistic test condition. For this verification test, no AL-PDU for control is multiplexed.
- **Multiplex Layer**
  The multiplex layer defined in H.223 Annex B was used. The optional header field was disabled.

The way to map the audio information into the MUX frame (MUX-PDU) is defined via a MultiplexEntryDescriptor in the MUX Table. In this verification test, one MUX-PDU contains only audio information. That is, the following MUX Table entry is pre-defined and used during the session:

- LCN1(audio), RC UCF

The audio channel is defined as segmentable to accommodate audio packets longer than the maximum length of MUX-SDU, and the MUX-SDU segmentation using the packet marker defined in H.223 is used. This implies that a part of the audio packet could be lost.

Multiplexer	H.223 Annex B (level 2)	
Audio Channel	AL Type	AL2
	Control Field (SN)	1 octet
	CRC	1 octet
	Retransmission	No
	Channel Type	segmentable
Control Channel	Not Used	
Multiplex Layer	H.223 Annex B with option	
	# Mux Table	3
	Table 1	{LCN1, RC UCF}[1]
	Flag	16 bit
	CT value	open
	Header Field	4 octet with optional field

Table 3-15: TransMux Layer Configuration

The error conditions of this test are described in the table here below. As a typical example of wireless mobile transmission channels, burst error channel is used as Physical Layer. Its error condition is defined as below:

Name	Average Bit Error Rate	Length of Burst Error
Critical Error Condition	10^{-3}	10 ms
Very Critical Error Condition	10^{-3}	1 ms

Table 3-16: Error Conditions

In the error conditions listed above, critical error condition corresponds to the point defined in the requirements (see [n2992]), and we can expect to prove that the error resilient audio encoder/decoder is compliant with the requirements for the error resilience as a result of the formal verification test. In actual wireless systems, the critical error condition corresponds to the worst cases that occur at the edge of radio service area, and the very critical condition is such bad condition that wouldn’t happen in an actual transmission channel in normal operation.

Error sequences were generated using software supplied by NTT DoCoMo [m2686]. Specifically, the Gilbert Model was used (a 2-state Markov Model). Bit errors occur only within the error burst, during which the bit error rate is 50 %. The probability of making a transition from a burst interval to a clear channel interval and back is:

Probability of BAD to GOOD (P_BADtoGOOD) = 1.0 / AverageBurstLength (in bits)
Probability of GOOD to BAD = AverageBER * P_BADtoGOOD * (0.5 - AverageBER)

---

[1] LCN: Audio

As the audio degradation by the errors depends on the error pattern, 25 kinds of error patterns are simulated in this test. The error pattern applied to the subjective evaluations will be automatically selected so that the produced SNR is the nearest one to the average SNR over all error patterns, so that the verification tests can give the most typical performance results.

It is assumed that AudioSpecificConfig() is transmitted through an error-free control channel.

4 Test Material

4.1 Test Items

Two selection panels have selected test items for session A1, A2, and A3. Whenever possible, the typical and critical test items and the training items were to be distributed among the four signal categories: speech, single instrument, music, and complex signals, as show in the following table:

	Speech	single instrument	music	Complex
Typical	1	1	1	1
Critical	1	1	1	1
Training	1	1	1	1

Table 4-1: Test item selection

4.2 Program Material Identified by Selection Panels

4.2.1 Session A1 – ER HILN

A selection panel at T-Nova has selected the test excerpts for session A1 and A2 (see [m5273]). These excerpts were selected from the set of 39 items used in the previous Audio on Internet tests (verification test for Version 1 tools, see [n2278], [n2425]).

No.	Item number	Name	Category
1	Item_07	Orchestral piece	Music
2	Item_11	We shall be happy	Single instrument
3	Item_12	Glockenspiel	Single instrument
4	Item_20	Percussion	Music
5	Item_29	Pop	Complex
6	Item_38	Erich Kaestner	Speech
7	Item_39	Complex sound + applause	Complex

Table 4-2: Test items for session A1 @ 6 kbit/s

No.	Item number	Name	Category
1	Item_13	Male German speech	Speech
2	Item_31	Classic	Complex
3	Item_37	Complex sound	Music

Table 4-3: Training items for session A1 @ 6 kbit/s

No.	Item number	Name	Category
1	Item_03	Castanets0	Single instrument
2	Item_04	Pitch pipe	Single instrument
3	Item_13	Male German speech	Speech
4	Item_15	Tracy Chapman	Complex
5	Item_18	Carmen	Music
6	Item_19	Accordion/Triangle	Music
7	Item_39	Complex sound + applause	Complex

Table 4-4: Test items for session A1 @ 16 kbit/s

No.	Item number	Name	Category
1	Item_14	Suzanne Vega	Music
2	Item_17	Haydn Trumpet Concert	Single instrument
3	Item_31	Classic	Complex

Table 4-5: Training items for session A1 @ 16 kbit/s

4.2.2 Session A2 – ER BSAC

As mentioned in the previous section, a panel at T-Nova has selected the test excerpts for session A2.

No.	Item number	Name	Category
1	Item_03	Castanets0	Single instrument
2	Item_04	Pitch pipe	Single instrument
3	Item_08	Contemporary pop music	Music
4	Item_13	Male German speech	Speech
5	Item_15	Tracy Chapmann	Complex
6	Item_18	Carmen	Music
7	Item_19	Accordion/Triangle	Music

Table 4-6: Test items for session A2 @ 64 kbit/s and above

No.	Item number	Name	Category
1	Item_14	Suzanne Vega	Music
2	Item_20	Percussion	Music
3	Item_39	Complex sound + applause	Complex

Table 4-7: Training items for session A2 @ 64 kbit/s and above

4.2.3 Session A3 – ER AAC LD

A selection panel at AT&T has selected the test excerpts for session A3 (see [m5012]). These excerpts were selected from the set of 51 items used in the NADIB tests (verification test for Version 1 tools, see [n2157], [n2276]).

No.	Item number	Name	Category
1	Item_02	Male English	Speech
2	Item_03	Male English	Speech
3	Item_18	Male English + music	Complex
4	Item_22	Bugpipe + drum	Music
5	Item_24	Piano	Single instrument
6	Item_36	Suzanne Vega	Music

Table 4-8: Test items for session A3 @ 64 kbit/s

No.	Item number	Name	Category
1	Item_11	Female English	Speech
2	Item_26	Male German + music	Complex

Table 4-9: Training items for session A3 @ 64 kbit/s

No.	Item number	Name	Category
1	Item_03	Male English	Speech
2	Item_05	Male French	Speech
3	Item_18	Male English + music	Complex
4	Item_24	Piano	Single Instrument
5	Item_31	Female/Male French	Speech
6	Item_36	Suzanne Vega	Music
7	Item_38	Vivaldi	Complex

Table 4-10: Test items for session A3 @ 32 kbit/s

No.	Item number	Name	Category
1	Item_11	Female English	Speech
2	Item_20	Female English + music	Complex
3	Item_29	Male French	Complex

Table 4-11: Training items for sessiion A3 @ 32 kbit/s

With respect to ER AAC LD 64 kbit/s @ 48 kHz (mono), the panel confirmed that AAC main 56 kbit/s is a sufficient reference codec.

With respect to ER AAC LD 32 kbit/s @ 48 kHz (mono), the panel confirmed that AAC main 24 kbit/s @ 24 kHz and G.722 64 kbit/s @ 16 kHz are sufficient reference codecs. CELP 24 kbit/s @ 16 kHz is a sufficient reference codec for the speech signals, however the panel observed that it may be too low being an anchor for music and complex (voice over music) signals.

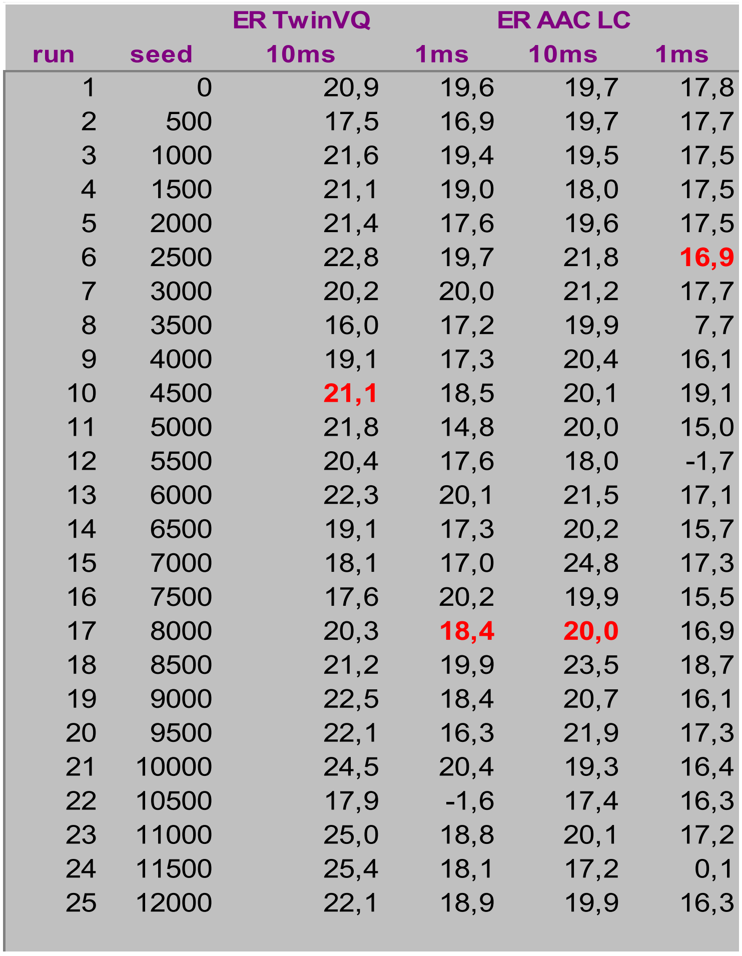
4.2.4 Session A4 – Error Robustness

Based on the test items used for the previous Audio on Internet tests (test D, see [n2278], [n2278]) the following 8 items are used:

No.	Item number	Category
1	01	speech
2	02	single instrument
3	11	single instrument
4	13	speech
5	20	complex
6	31	classical
7	33	complex
8	37	pop

Table 4-12: Items used for session A4

For session A4, NTT DoCoMo has performed a selection of a typical error pattern based on objective measurement:



		ER TwinVQ		ER AAC LC	
run	seed	10ms	1ms	10ms	1ms
1	0	20,9	19,6	19,7	17,8
2	500	17,5	16,9	19,7	17,7
3	1000	21,6	19,4	19,5	17,5
4	1500	21,1	19,0	18,0	17,5
5	2000	21,4	17,6	19,6	17,5
6	2500	22,8	19,7	21,8	**16,9**
7	3000	20,2	20,0	21,2	17,7
8	3500	16,0	17,2	19,9	7,7
9	4000	19,1	17,3	20,4	16,1
10	4500	**21,1**	18,5	20,1	19,1
11	5000	21,8	14,8	20,0	15,0
12	5500	20,4	17,6	18,0	-1,7
13	6000	22,3	20,1	21,5	17,1
14	6500	19,1	17,3	20,2	15,7
15	7000	18,1	17,0	24,8	17,3
16	7500	17,6	20,2	19,9	15,5
17	8000	20,3	**18,4**	**20,0**	16,9
18	8500	21,2	19,9	23,5	18,7
19	9000	22,5	18,4	20,7	16,1
20	9500	22,1	16,3	21,9	17,3
21	10000	24,5	20,4	19,3	16,4
22	10500	17,9	-1,6	17,4	16,3
23	11000	25,0	18,8	20,1	17,2
24	11500	25,4	18,1	17,2	0,1
25	12000	22,1	18,9	19,9	16,3

Table 4-13: SNR values for items in session A4, selected items are bold

5 Test Methodology

5.1 Test Method and Test Design for Sessions A1, A2, and A3

The subjective assessment of sound quality was done according to ITU-Recommendation BS.1284 [bs1284]. This was chosen to permit these results to be compared to those of the MPEG-4 Version 1 tests.

The following 5-grade scale was used:

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 5-1: BS.1284 Quality scale

In order to achieve higher precision in the test results the quality scale was used as a continuous scale with one decimal place.

The listening test was designed as follows:

- training with the corresponding selected training items
- stimuli presentation in pairs A-B (called a trial), with 'A' always the reference stimulus and 'B' the processed version
- Each grading phase was divided into sections of approx. 20 minutes length.

5.1.1 Specifics of Session A1

- 16 subjects participated; none of the subjects has reported having hearing impairments
- There was an acoustical announcement of the test item number
- Headphones were STAX LAMBDA NOVA
- Presentation was done via DAT
- There were 2 listeners at a time
- A separate grading phase took place for session part A1 @ 6 kbit/s, and session part A1 @ 16 kbit/s

5.1.2 Specifics of Sessions A2 and A3

- 24 subjects participated; none of the subjects reported having hearing impairments; all were from 20 to 30 years of age, most of the listeners were students at a music academy
- Headphones were STAX LAMBDA NOVA
- PC based presentation was used. All test items were upsampled to 48 kHz using the default setting of the ResampAudio tool from the AFsp library.
- Presentation order and item numbers were shown on a small display synchronized with the playback, but the scores were recorded on paper.
- There were four listeners at a time, using a common randomized presentation order. Thus each test has 6 different randomized sequences, since the number of subjects was 24.
- Separate grading phases took place for session A2, session part A3 @ 64 kbit/s, and session part A3 @ 32 kbit/s; listening was done in this order from morning to afternoon.

5.2 Test Method and Test Design for Session A4

"Subjective assessment of sound quality" (MUSHRA) [included in n2953] was the test method used in Session A4. (This method is a proposed standard at EBU and ITU-R.)

Session A4 was separated into two parts, each with a common channel bit rate and common number of signal channels, designated as follows:

A4 @ 16 kbit/s	ER TwinVQ	16 kbit/s, excluding EP or ER tool rate, mono stimuli
A4 @ 96 kbit/s	ER AAC LC	96 kbit/s, excluding EP or ER tool rate, stereo stimuli

Each of A4 @ 16 kbit/s and A4 @ 96 kbit/s were conducted at two test locations: NTT DoCoMo and FhG. Each test at each location had sufficient listeners to be evaluated on its own, and the results will be reported separately in this report. The motivation for the duplicate testing was that since each of these laboratories is a proponent of the technology, each result could serve as a crosscheck on the other.

The following stimuli were used as references:

1. full bandwidth hidden reference
2. low pass filtered hidden reference (7 kHz)
3. low pass filtered hidden reference (3.5 kHz)

The following additional reference stimuli was added for A4 @ 16 kbit/s only:

4. low pass filtered hidden reference (1.7 kHz)

The following stimuli (either ER TwinVQ ER for A4 @ 16 kbit/s or ER AAC LC for A4 @ 96 kbit/s) were used as test stimuli:

5. undistorted (clear channel condition)
6. distorted (critical channel condition)
7. distorted (very critical channel condition)

A4 @ 16 kbit/s had a total of 7 stimuli, and A4 @ 96 kbit/s had a total of 6 stimuli.

The number of listeners in each test at each test site was as follows:

A4 @ 16 kbit/s	FhG	NTT DoCoMo	total
listeners	27	18	45
expert	17		
non-exert	10		

A4 @ 96 kbit/s	FhG	NTT DoCoMo	total
listeners	27	20	47
expert	17		
non-expert	10		

The parameters of the listening test design were as follows:

- Stimuli presentation was not fixed, but rather the test subject had the possibility to switch between all instances of the audio signal in any order as often as he or she desired.
- Headphones were STAX (preferred STAX LAMBDA PRO)
- There was one listener at a time, due to computer based grading procedure.
- Audio was presented via computer-control.
- Grading was performed via computer-control

5.3 Training of Subjects

Prior to the sessions, all subjects in all tests (A1, A2, A3, and A4) participated in a training session. The training sessions encompassed the following:

- For session A1, there was training at both bit rates with respect to the codec under test (6 kbit/s and 16 kbit/s).
- For session A2, there was training at the lowest and at the highest bit rate with respect to the codec under test (64 kbit/s and 96 kbit/s).
- For session A3, there was training at both bit rates with respect to the codec under test (32 kbit/s and 64 kbit/s).
- For session A4, there was training for both bit rates with respect to the codec under test (16 kbit/s and 96 kbit/s).
- If several reference signals were used within a session, all of them were used in training.

The first step of training is to listen to the training items in order to become familiar with the nature of the artifacts. The subjects can discuss the perceived artifacts, but subjects are not allowed to talk about specific grades in order to avoid bias in individual grading. The randomization of the order of presentation of the training items and the number of repetitions of the items was at the discretion of each listening test site.

The second step of the training is to run a dummy grading of the training items using the grading facility (paper sheet or on-screen display) to become familiar with this tool for the subsequent grading phase.

The goal of the training is to make the subjects familiar what to listen to and how to grade. For test session A4, instructions stated in Annex C.4.1 have been given to the listener.

6 Test Results

6.1 *Session A1 – ER HILN*

6.1.1 Analysis Method

After the subjective listening tests were completed, average scores and 95 % confidential intervals were calculated for selected pooling of the data. Specifically, pooling of data was done as follows:

Result	Pooling of data
For each system (Overall Results)	All listeners for all test items for that system
For each item and each system (Codec-by-Codec Results, Item-by-Item Results)	All listeners for that test item and that system

Table 6-1: Pooling of data

In this table "system" refers to a codec at a specific bit rate. The second pooling of the data is presented twice, first in the "Codec-by-Codec Results" section as one plot for each test item, and then in the "Item-by-Item Results" section as one plots for each system. Data from all listeners were used in the analysis.

6.1.2 Results

6.1.2.1 Overall Results

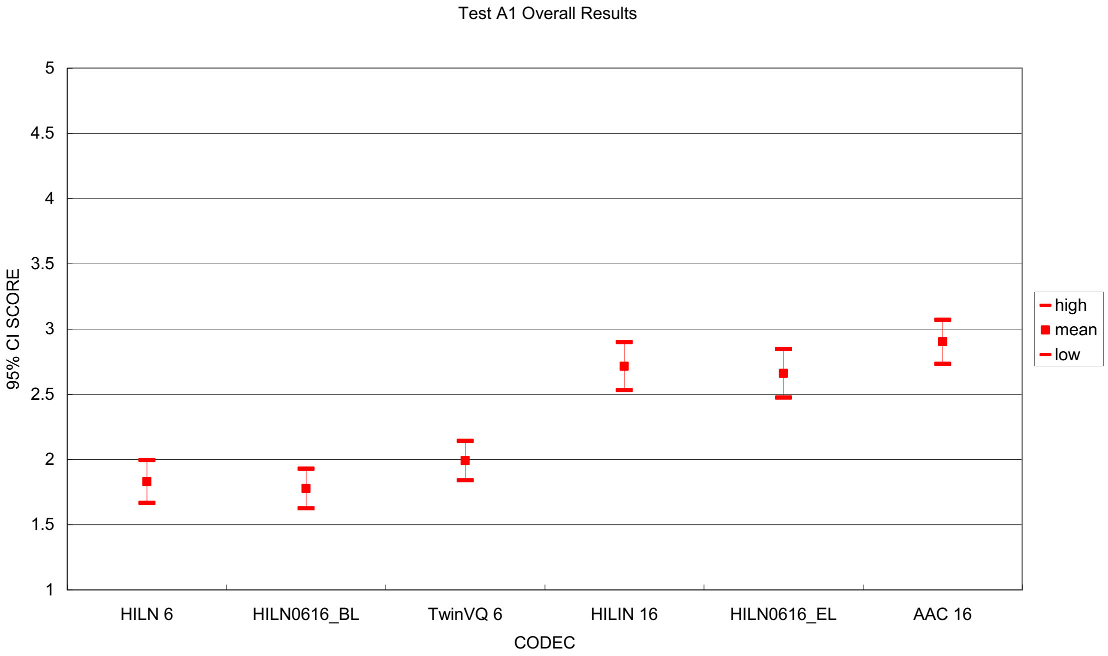


Figure 6-1: Session A1 Overall Results

6.1.2.2 *Codec-by-Codec Results: 6 kbits/s*

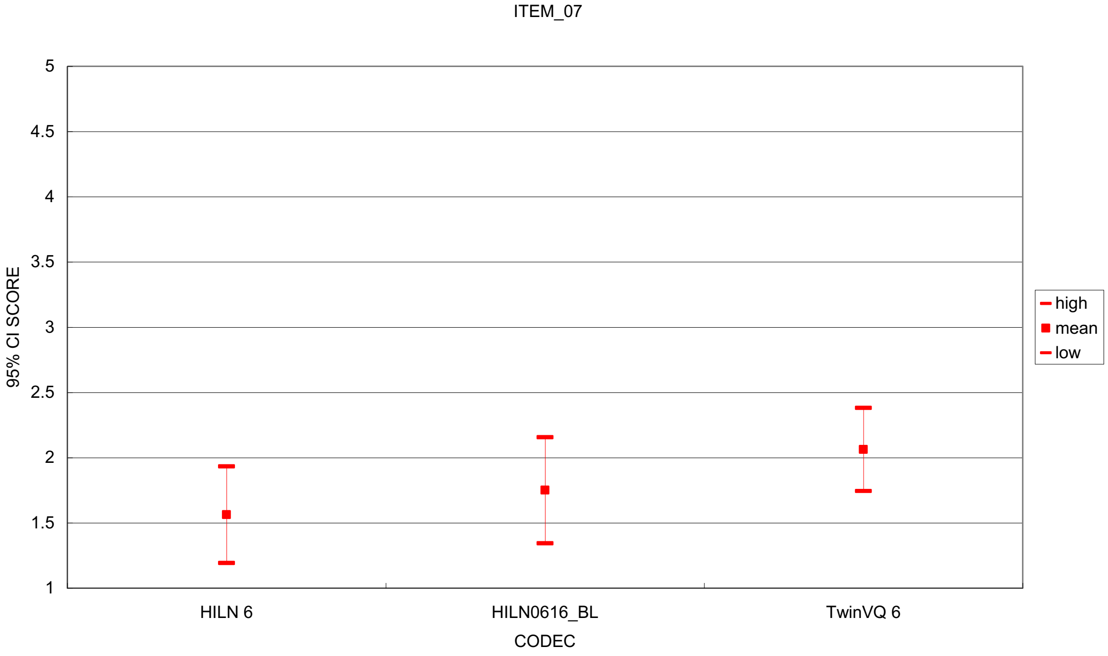


Figure 6-2: Item 07 (Orchestral Piece: Music)

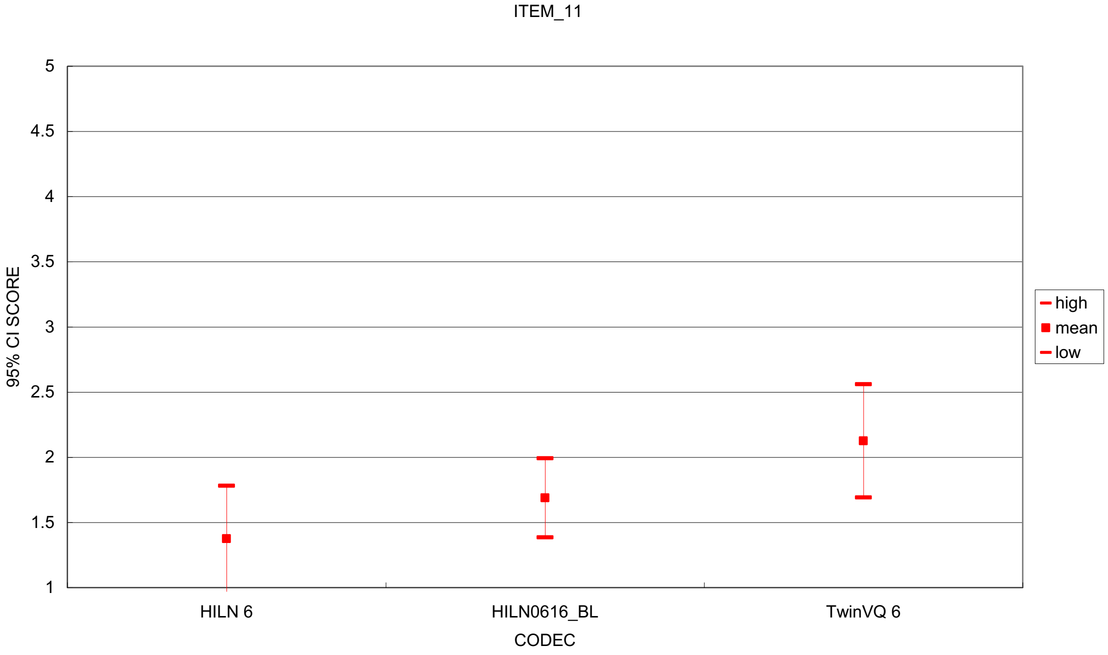


Figure 6-3: Item 11 (We shall be happy: Single instrument)

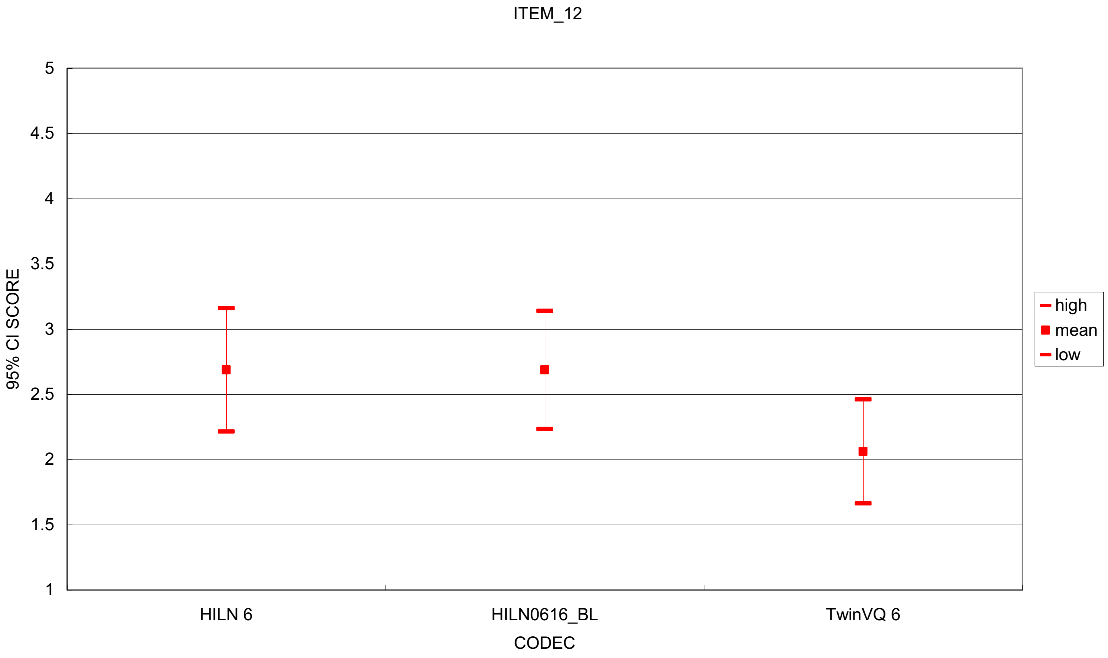


Figure 6-4: Item 12 (Glockenspiel: Single instrument)

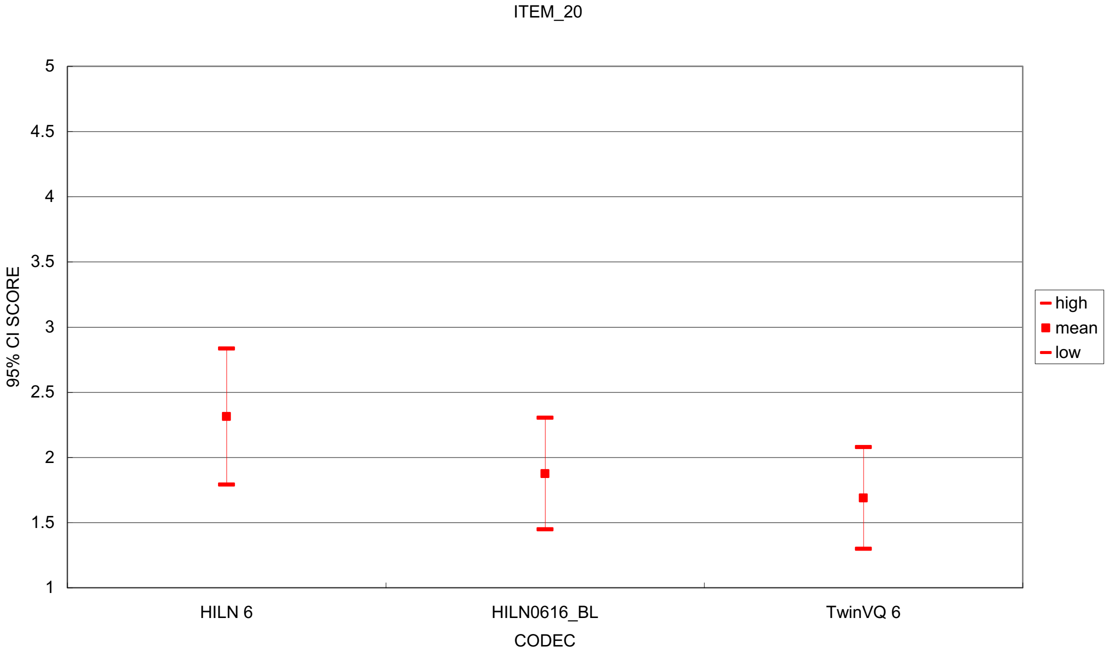


Figure 6-5: Item 20 (Percussion: Music)

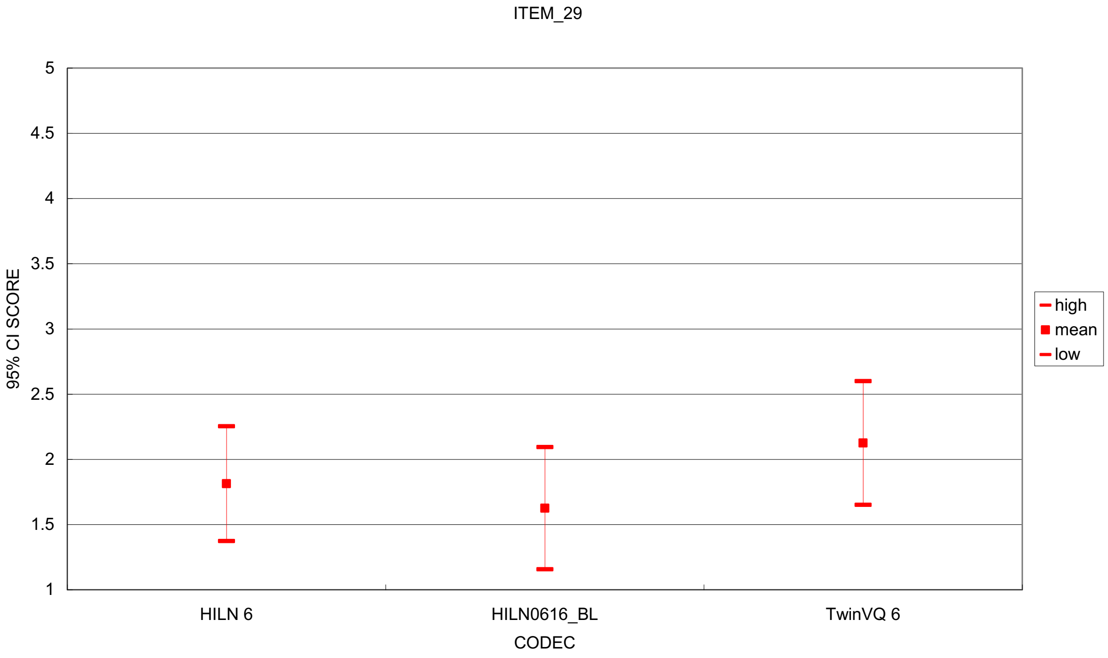


Figure 6-6: Item 29 (Pop: Complex)

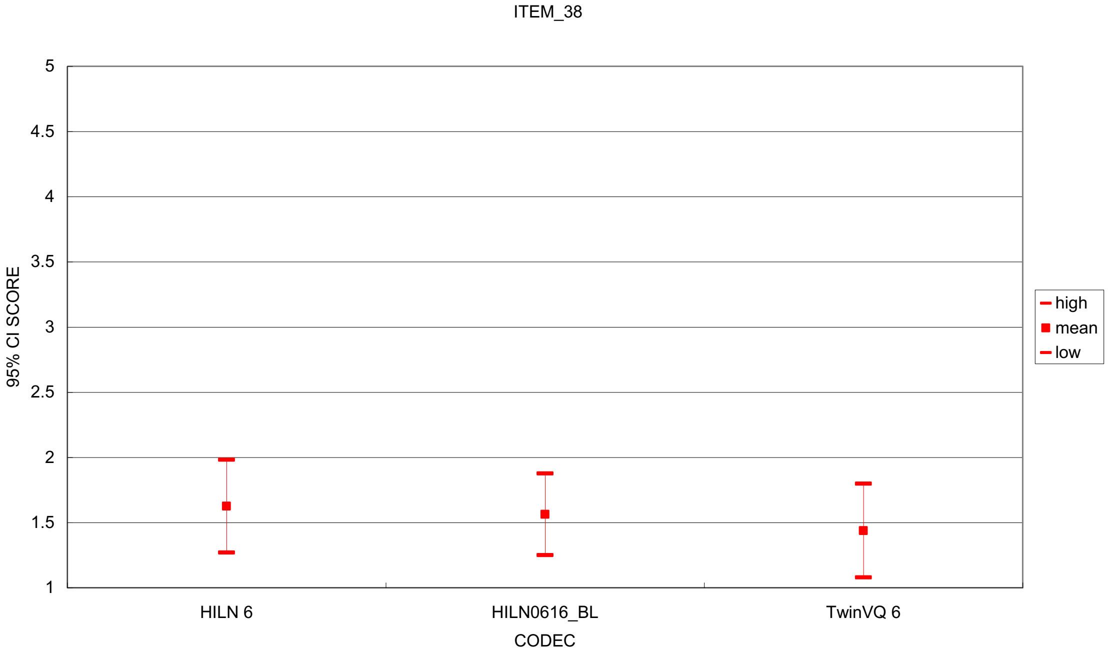


Figure 6-7: Item 38 (Erich Kaestner: Speech)

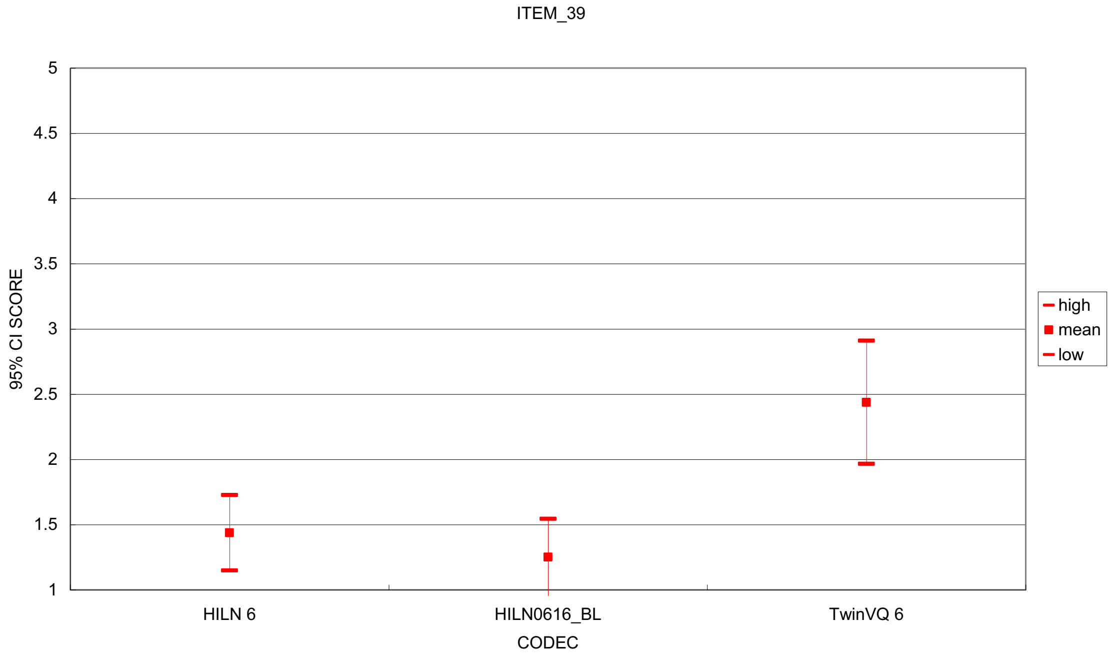


Figure 6-8: Item 39 (Complex sound + applause: Complex)

6.1.2.3 Codec-by-Codec Results: 16 kbits/s

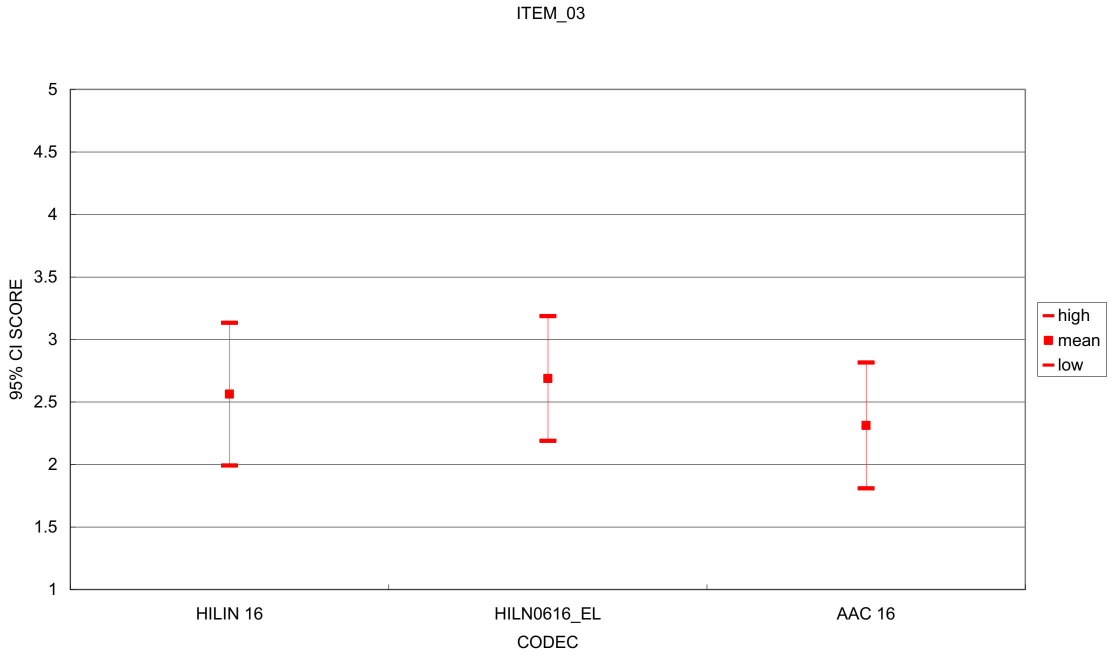


Figure 6-9: Item 03 (Castanets: Single instrument)

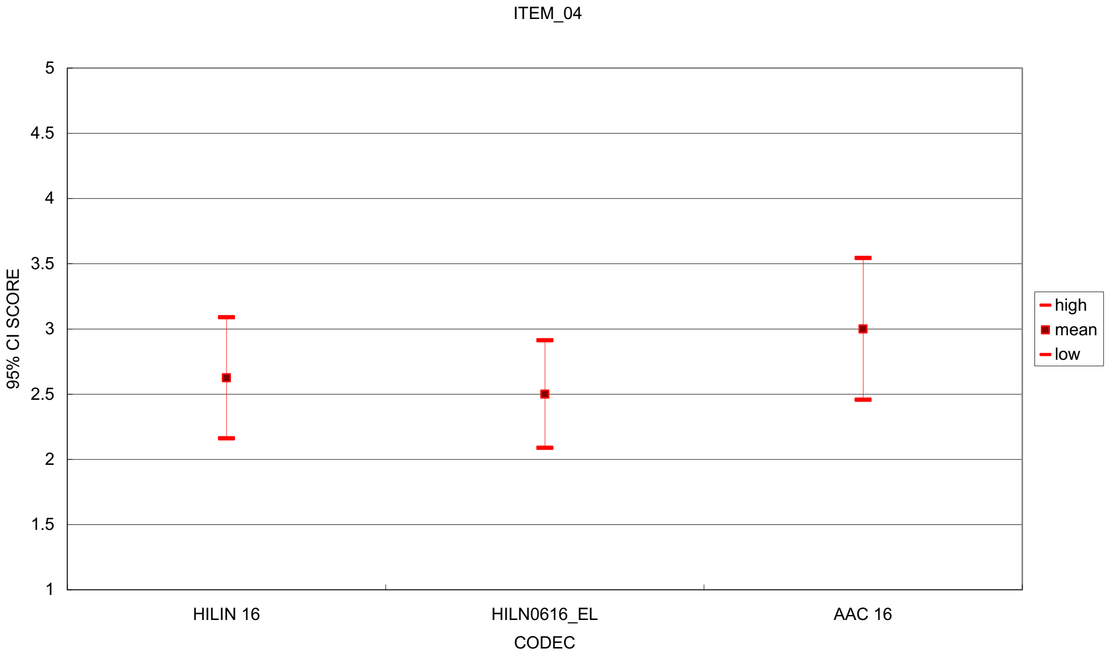


Figure 6-10: Item 04 (Pitch pipe: Single instrument)

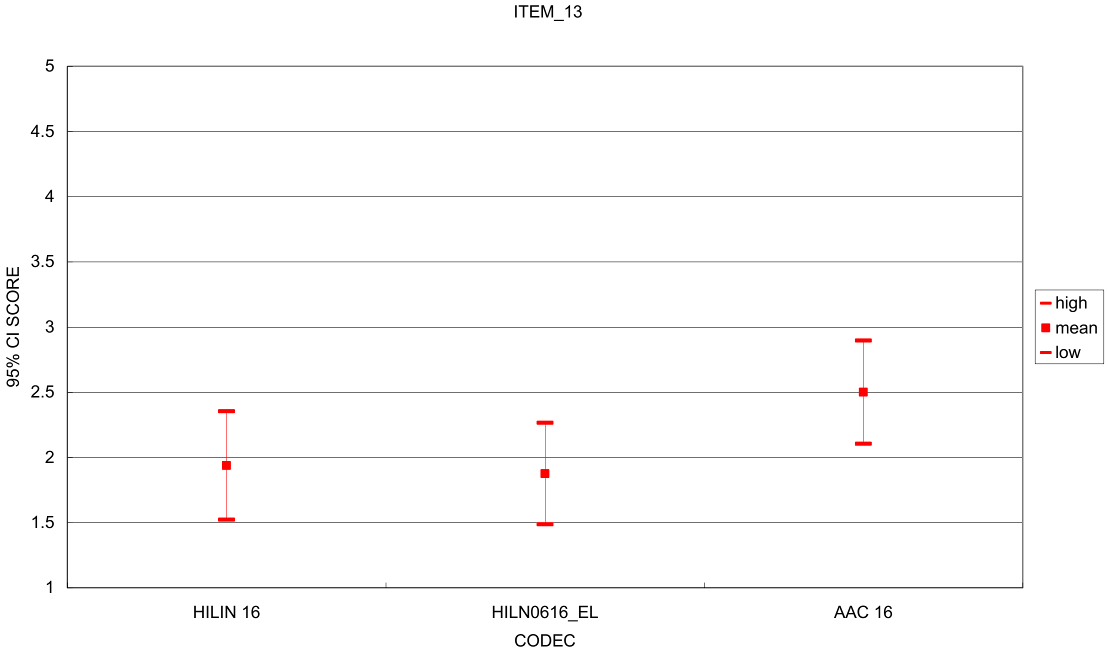


Figure 6-11: Item 13 (Male German Speech: Speech)

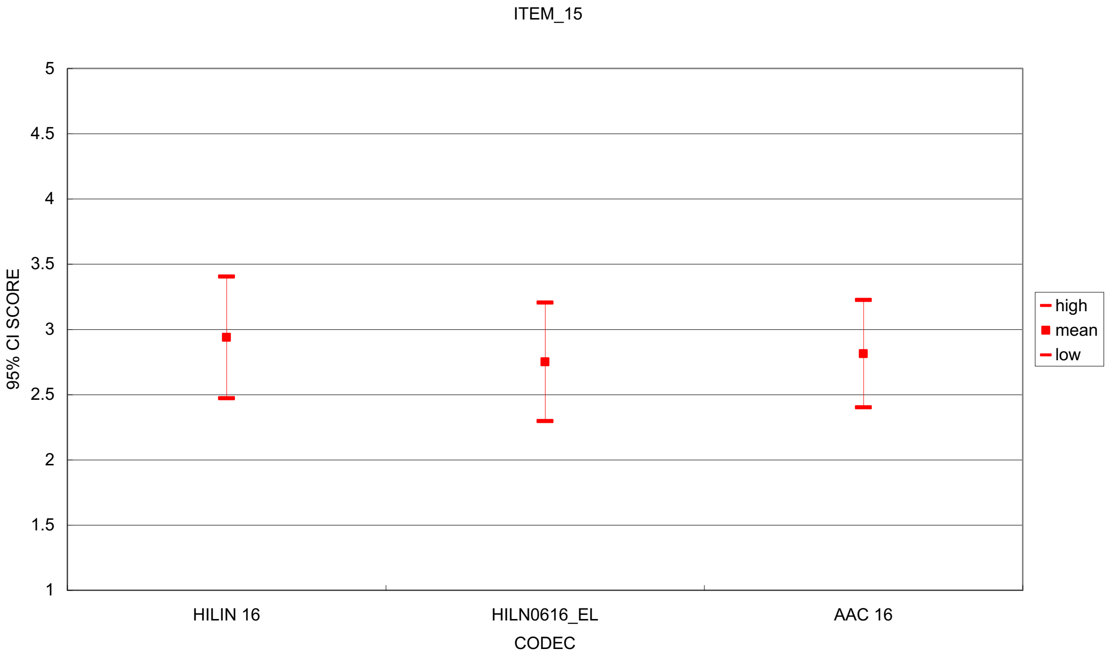


Figure 6-12: Item 15 (Tracy Chapman: Complex)

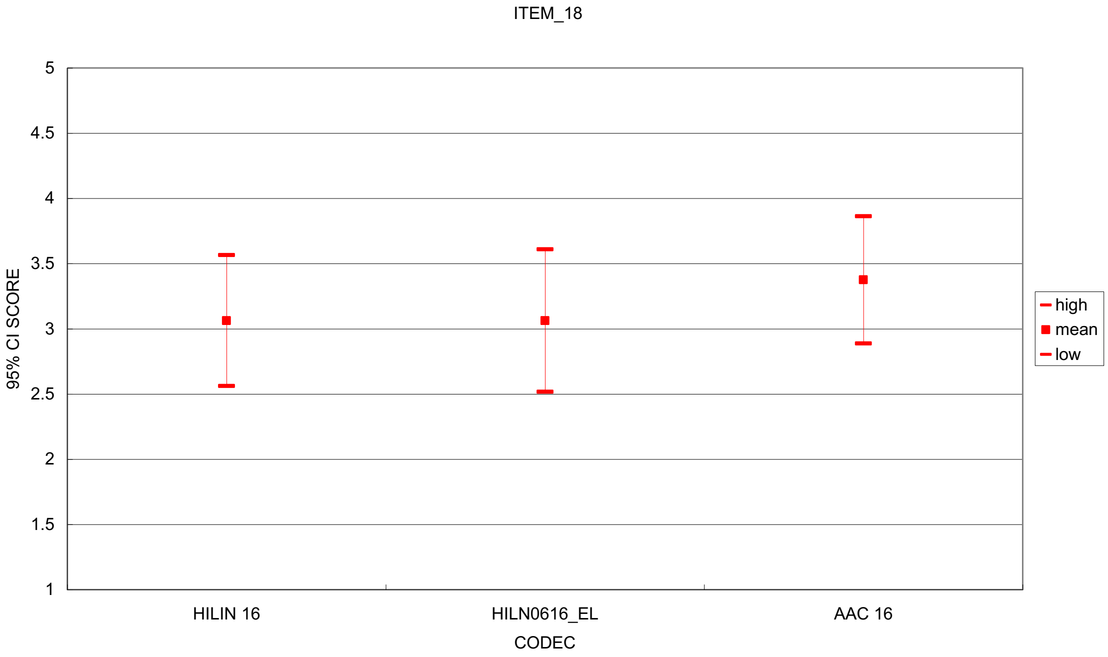


Figure 6-13: Item 18 (Carmen: Music)

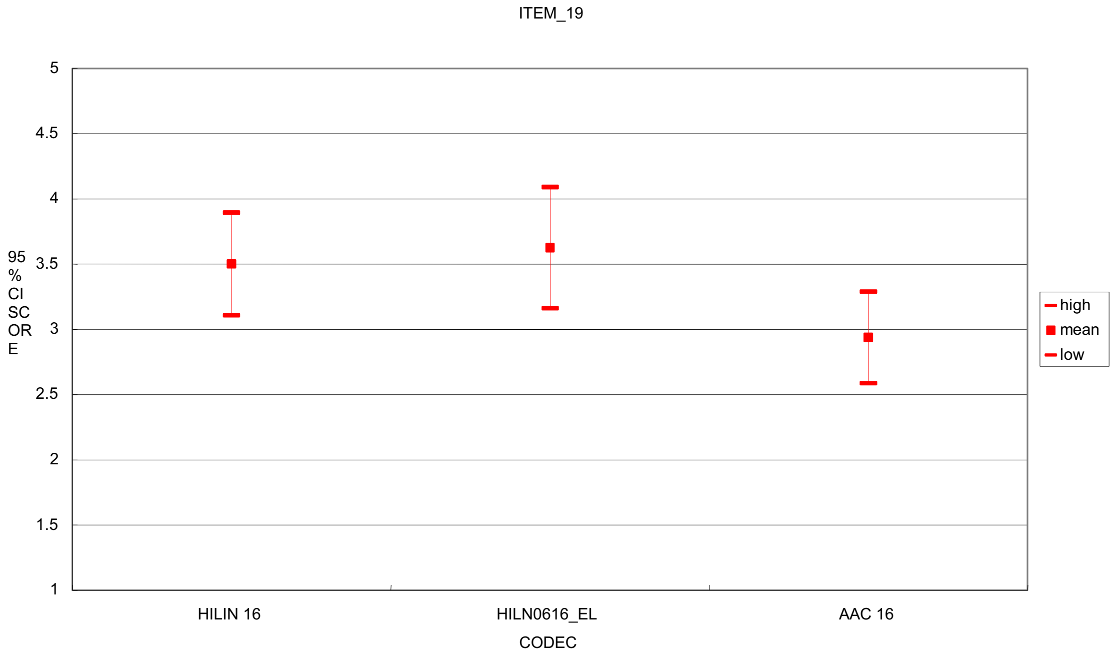


Figure 6-14: Item 19 (According/Triangle: Music)

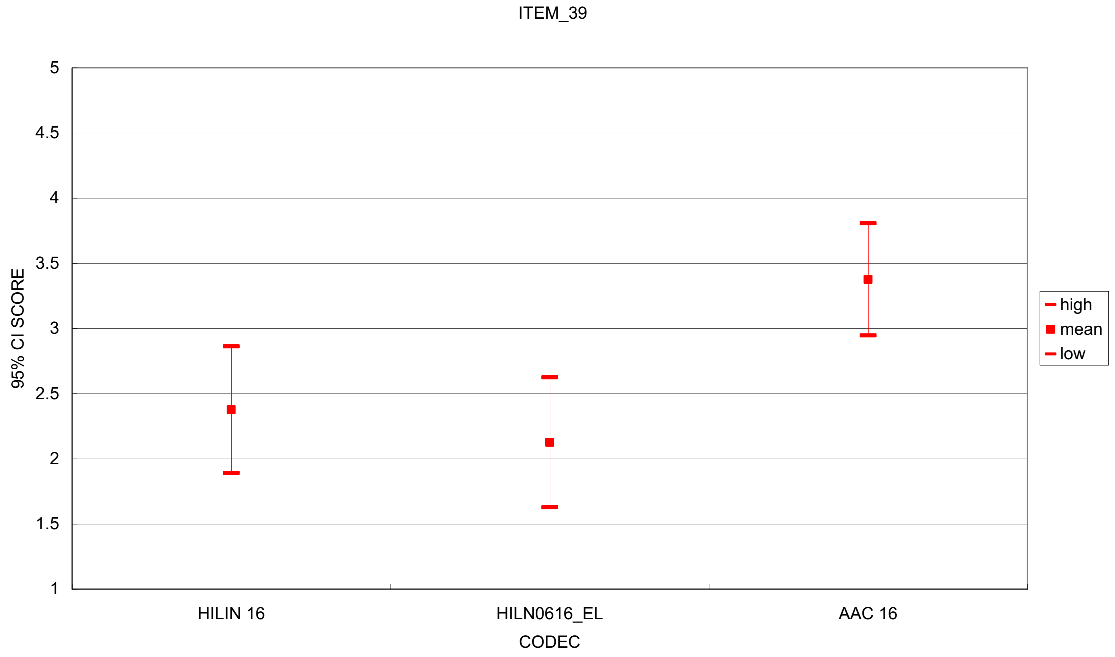


Figure 6-15: Item 39 (Complex sound + applause: Complex)

6.1.2.4 Item-by-Item Results

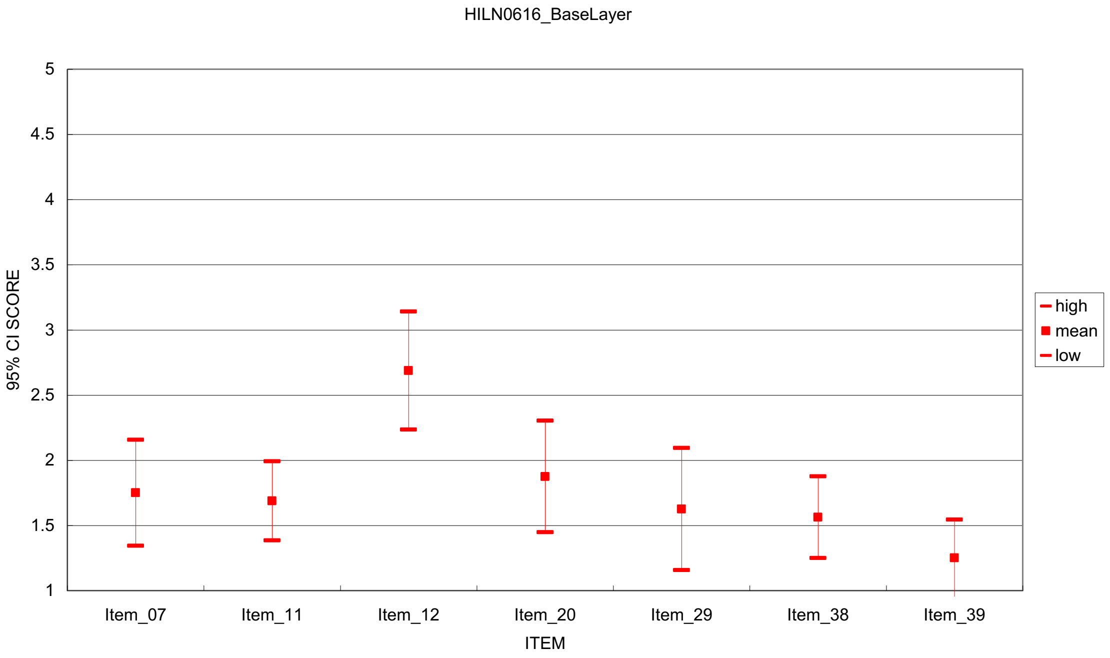


Figure 6-16: ER HILN 6 kit/s @ 16 kHz (mono)

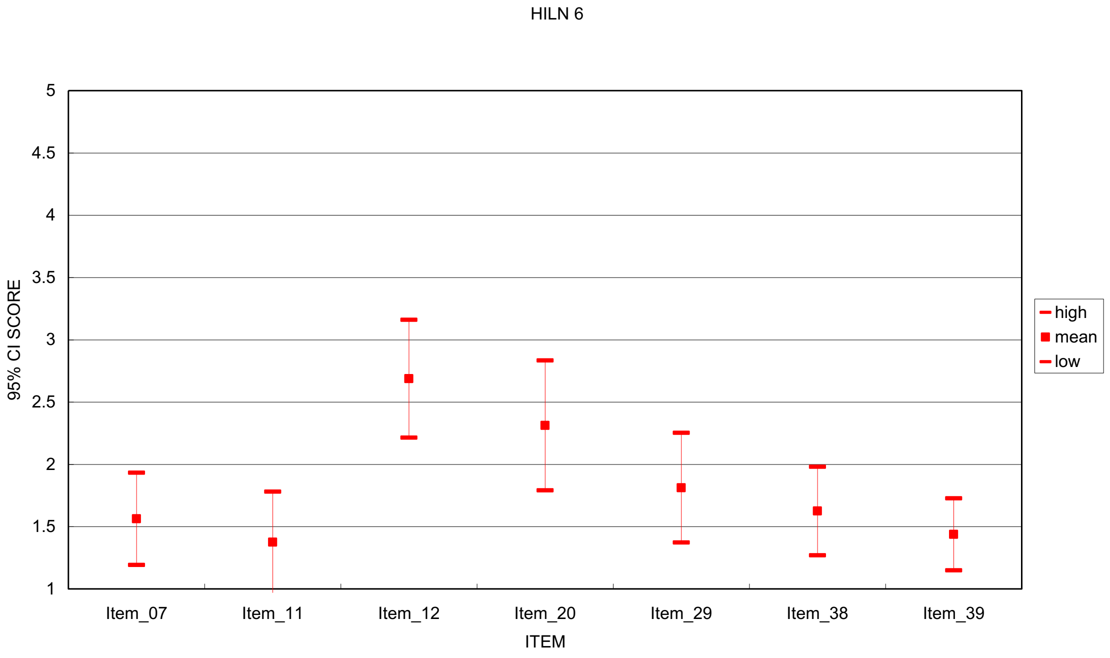


Figure 6-17: ER HILN 6 kit/s @ 16 kHz (mono) based on scalable configuration (6 kbit/s + 10 kbit/s)

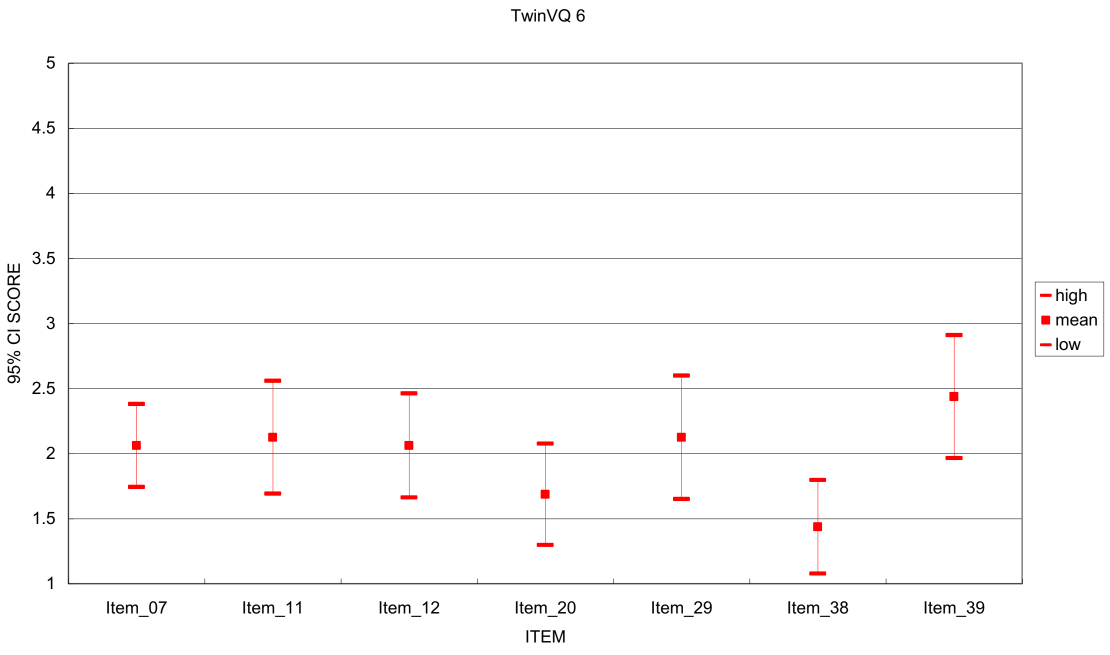


Figure 6-18: TwinVQ 6 kit/s @ 16 kHz (mono)

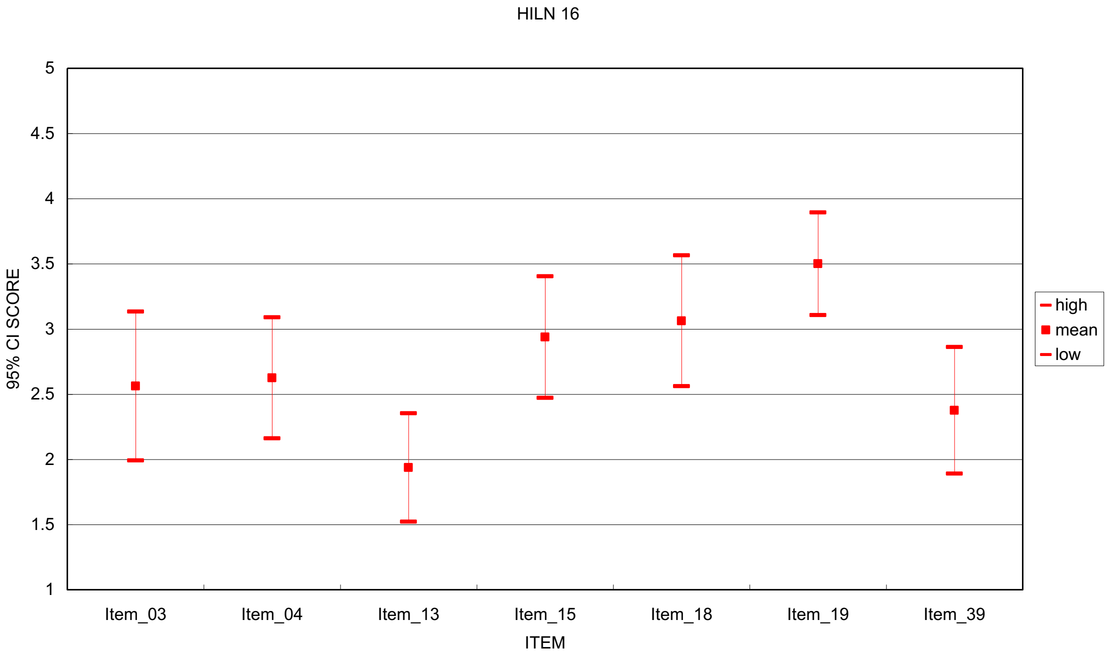


Figure 6-19: ER HILN 16 kit/s @ 16 kHz (mono)

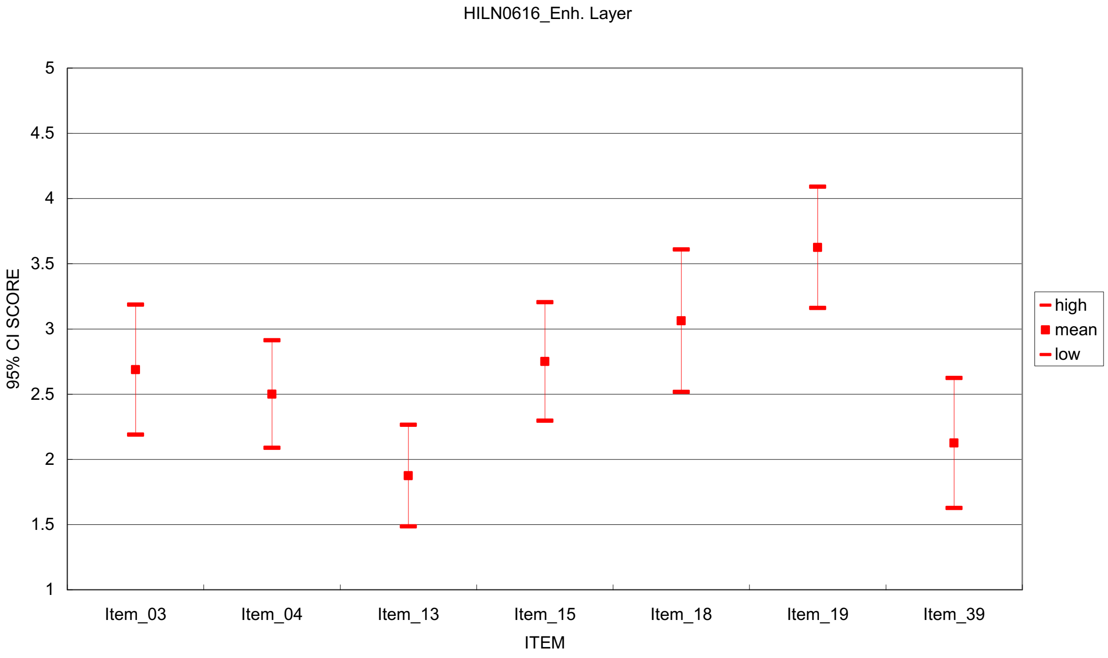


Figure 6-20: ER HILN 16 kit/s @ 16 kHz (mono) based on scalable configuration (6 kbit/s + 10 kbit/s)

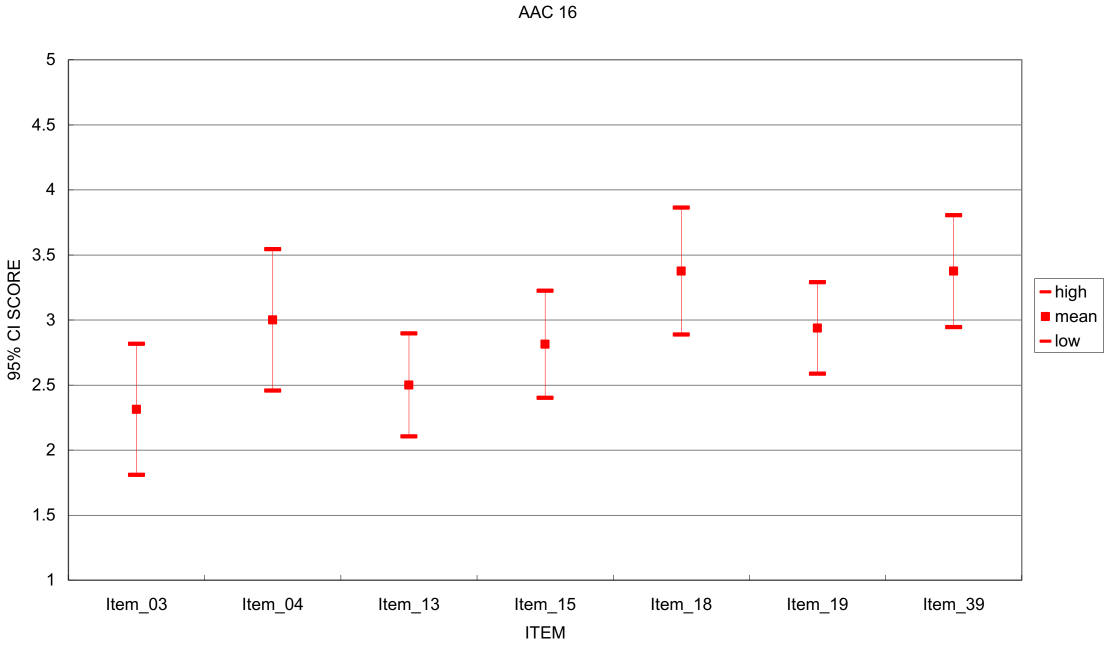


Figure 6-21: AAC main 16 kit/s @ 22.05 kHz (mono)

6.1.3 Discussion

6.1.3.1 Overall Results

The following statements are valid based on the mean scores and associated two-sided 95 % confidence intervals:

- At 6 kbit/s no codec was statistically different from any other.
- At 16 kbit/s no codec was statistically different from any other.

Therefore one can conclude that

- At 6 kbit/s, the bit rate scalability feature of ER HILN base plus enhancement layer coding (6 kbit/s +10 kbit/s) does not incur any penalty in quality relative to non-scalable ER HILN at 6 kbit/s.
- Similarly, at 16 kbit/s, the bit rate scalability feature of ER HILN base plus enhancement layer coding (6 kbit/s + 10 kbit/s) does not incur any penalty in quality relative to non-scalable ER HILN at 16 kbit/s.
- ER HILN at both 6 kbit/s and 16 kbit/s has performance comparable to other MPEG-4 coding technology operating at similar bit rates, but provides the additional capability of independent audio signal speed or pitch change while decoding.

6.1.3.2 Codec-by-Codec Results

In the following tables, the first column indicates a system (codec at a specified bit rate) and the second column associates a number with that system. The numbers, indicating systems, appear again as column headings over the body of the table. In the body of the table, the numeric entries indicate for how many test items the performance of the system in that row is statistically better than the performance of the system in that column. In this test there were a total of 7 test items.

Codec	No.	1	2	3
ER HILN 6 kbit/s	1		0	0
ER HILN BL 6 kbit/s	2	0		0
TwinVQ 6 kbit/s	3	1	1	

Table 6-2: Number of items with statistically significant differences

Codec	No.	1	2	3
ER HILN 16 kbit/s	1		0	0
ER HILN EL 16 kbit/s	2	0		0
AAC main 16 kbit/s	3	1	1	

Table 6-3: Number of items with statistically significant differences

6.1.3.3 Comparison with earlier Test Results

In the "Audio On Internet" verification test [n2425], conducted in Summer 1998, an earlier version of the ER HILN coder was assessed. At 6 kbit/s and 16 kbit/s, it showed a significantly worse overall performance than TwinVQ at 6 kbit/s and AAC main at 16 kbit/s, respectively. The quality of this earlier ER HILN was highly dependent on the test material, and for some items it even was better than TwinVQ or AAC main. Therefore, it was concluded to continue work on ER HILN in Version 2 to improve its coding quality, especially for critical test material. The results of the Version 2 verification test session A1 show that the subjective quality of ER HILN is significantly improved in Version 2 and now is comparable to other MPEG-4 coding technology operating at similar bit rates.

6.2 Session A2 – ER BSAC

6.2.1 Analysis Method

After the subjective listening tests were completed, average scores and 95 % confidential intervals were calculated for selected pooling of the data. Specifically, pooling of data was done as follows:

Result	Pooling of data
For each system (Overall Results)	All listeners for all test items for that system
For each item and each system (Codec-by-Codec Results, Item-by-Item Results)	All listeners for that test item and that system

Table 6-4: Pooling of data

In this table “system” refers to a codec at a specific bit rate. The second pooling of the data is presented twice, first in the “Codec-by-Codec Results” section as one plot for each test item, and then in the “Item-by-Item Results” section as one plots for each system. Data from all listeners were used in the analysis.

6.2.2 Results

6.2.2.1 Overall Results

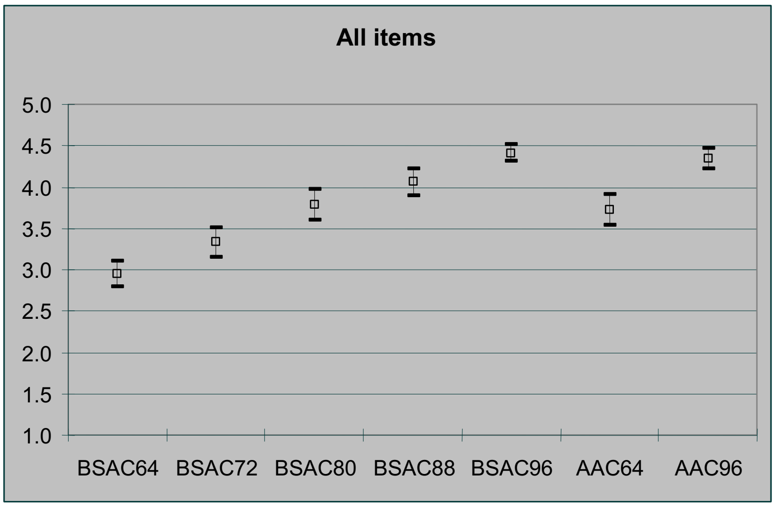


Figure 6-22: Average scores for all items

6.2.2.2 Codec-by-Codec Results

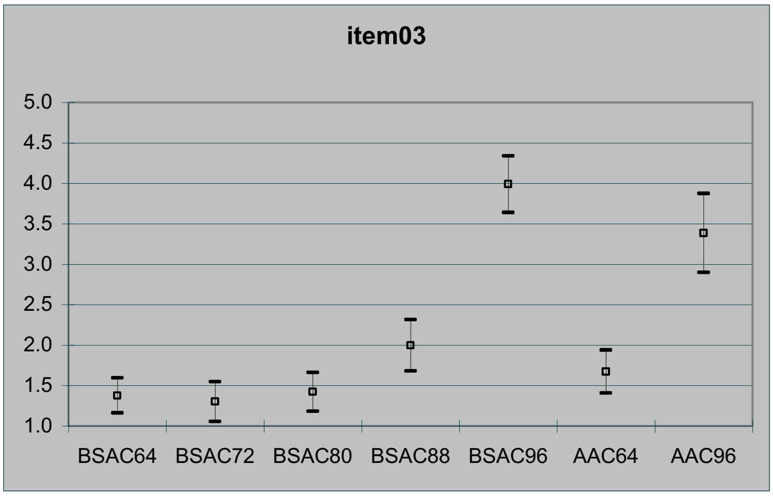


Figure 6-23: Item by item scores (item03)

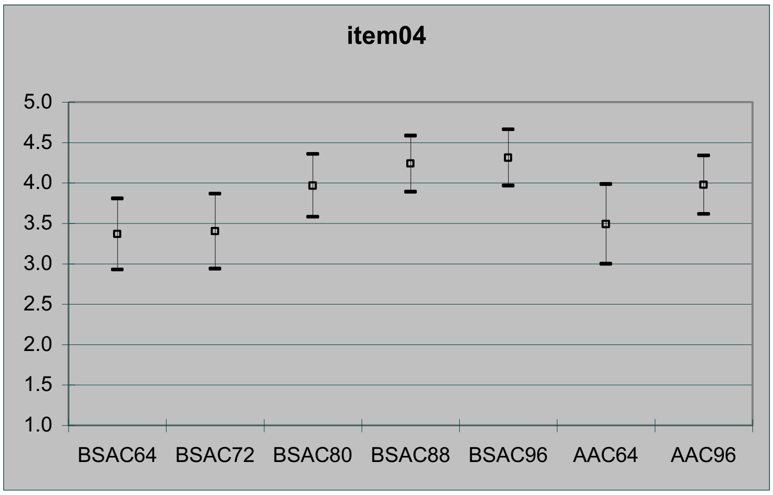


Figure 6-24: Item by item scores (item04)

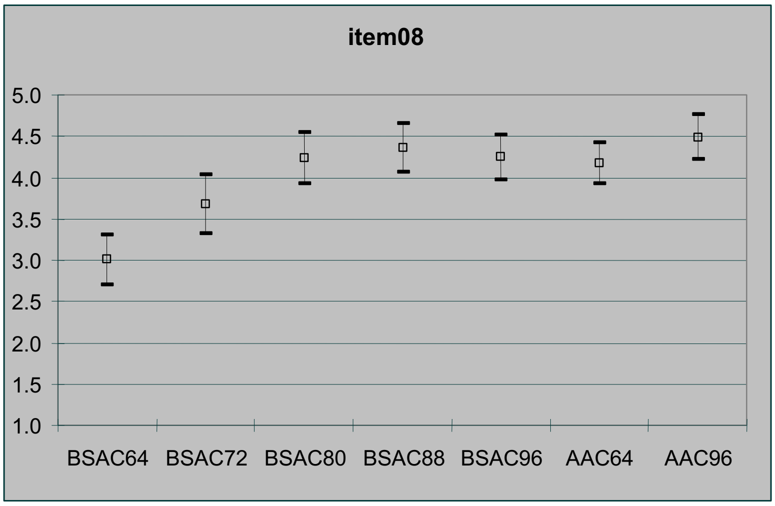


Figure 6-25: Item by item scores (item08)

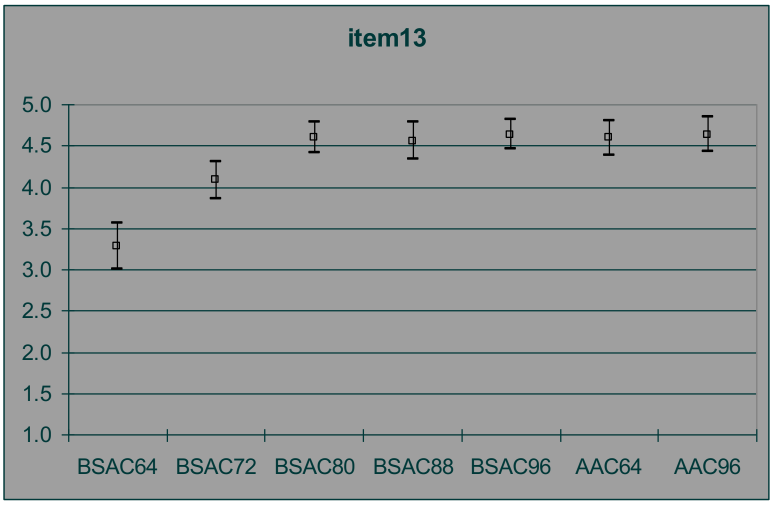


Figure 6-26: Item by item scores (item13)

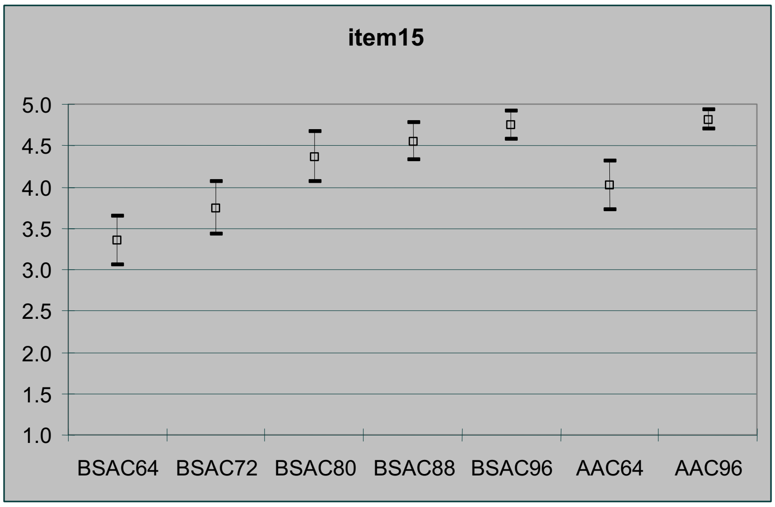


Figure 6-27: Item by item scores (item15)

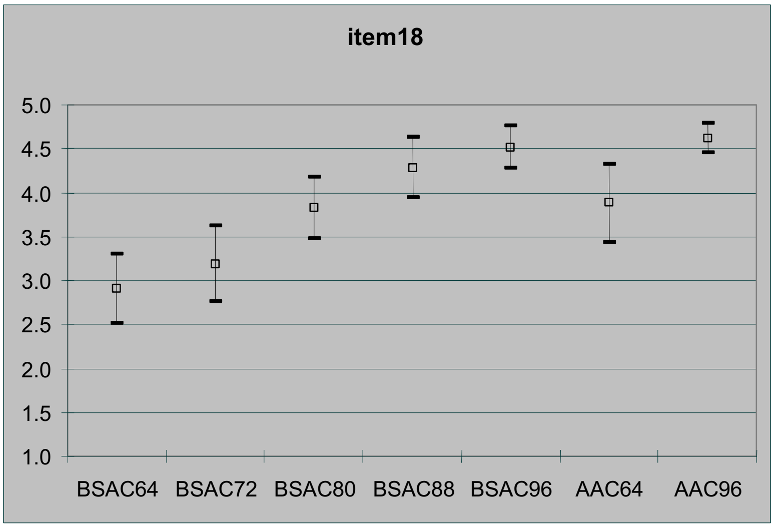


Figure 6-28: Item by item scores (item18)

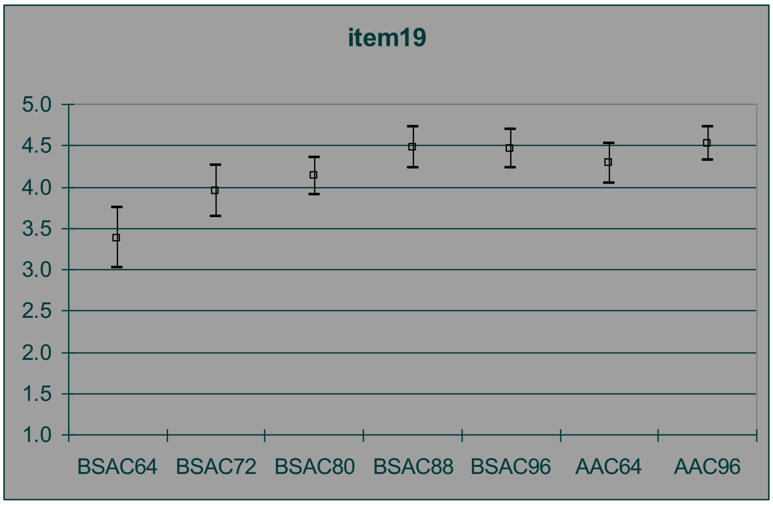


Figure 6-29: Item by item scores (item19)

6.2.2.3 Item-by-Item Results

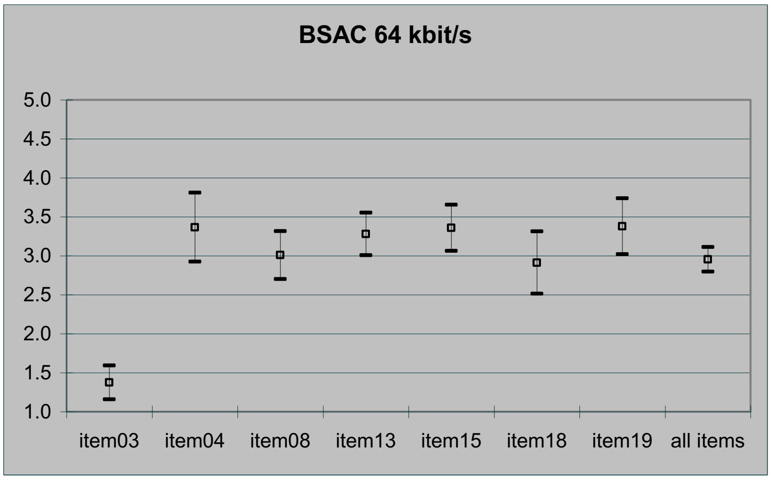


Figure 6-30: Scores for ER BSAC 64 kbit/s

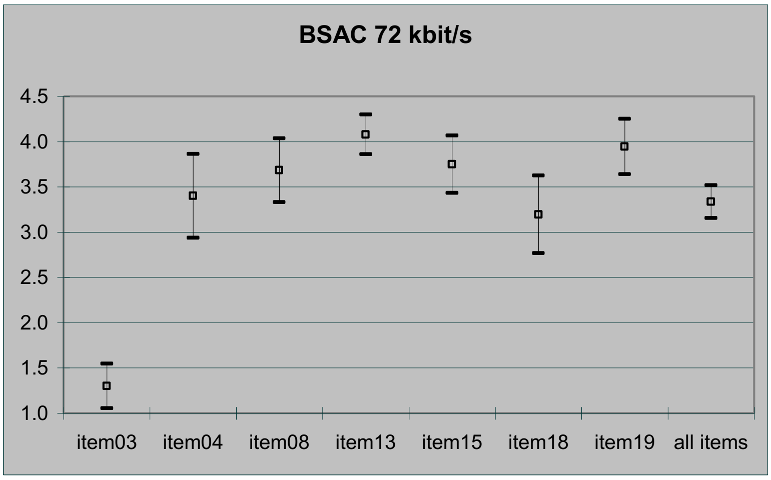


Figure 6-31: Scores for ER BSAC 72 kbit/s

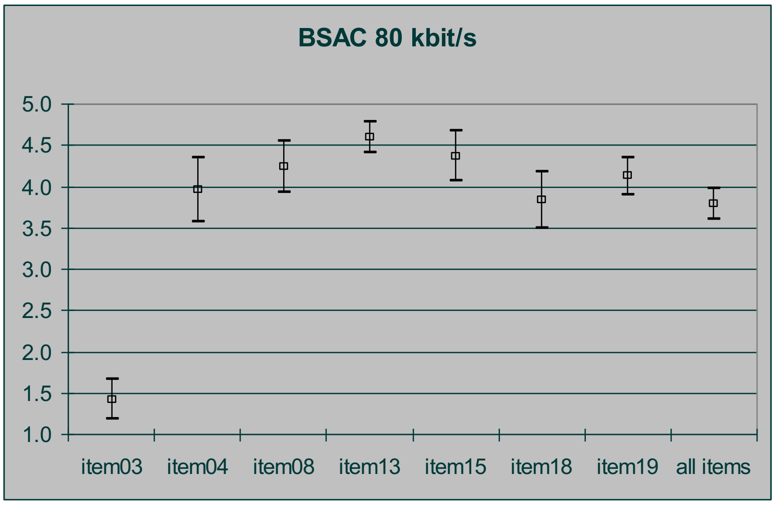


Figure 6-32: Scores for ER BSAC 80 kbit/s

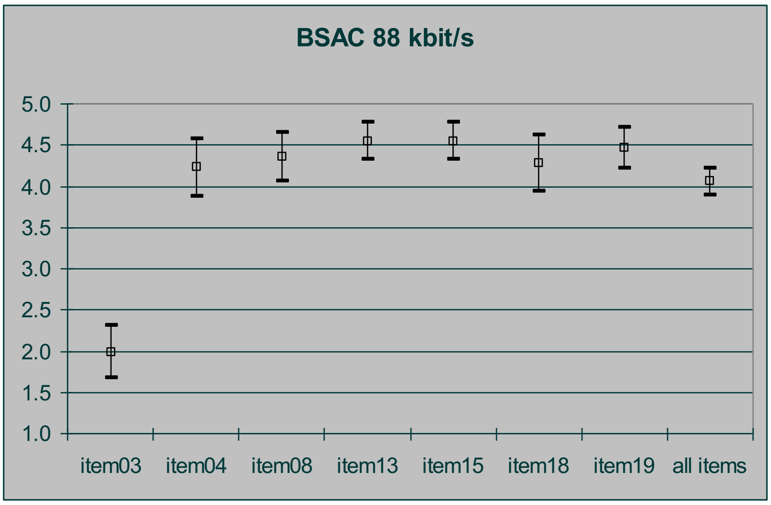


Figure 6-33: Scores for ER BSAC 88 kbit/s

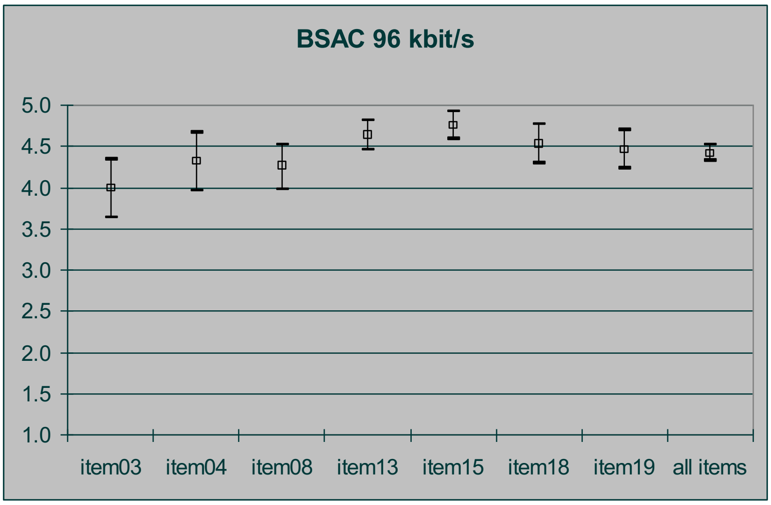


Figure 6-34: Scores for ER BSAC 96 kbit/s

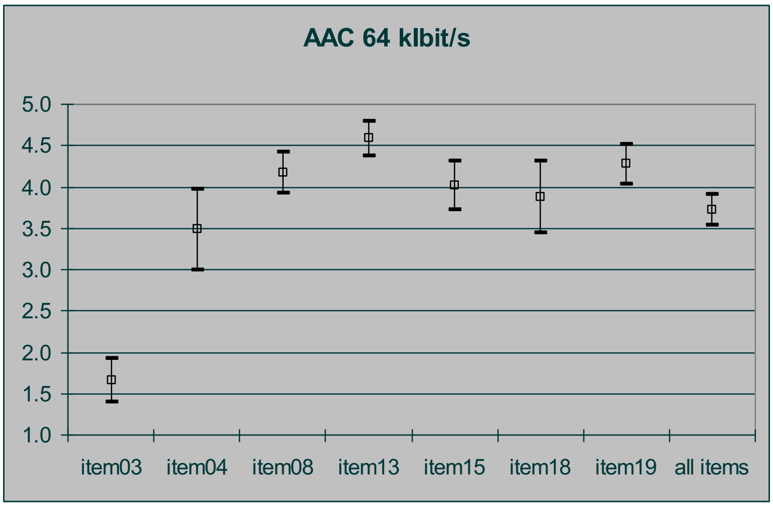


Figure 6-35: Scores for AAC 64 kbit/s

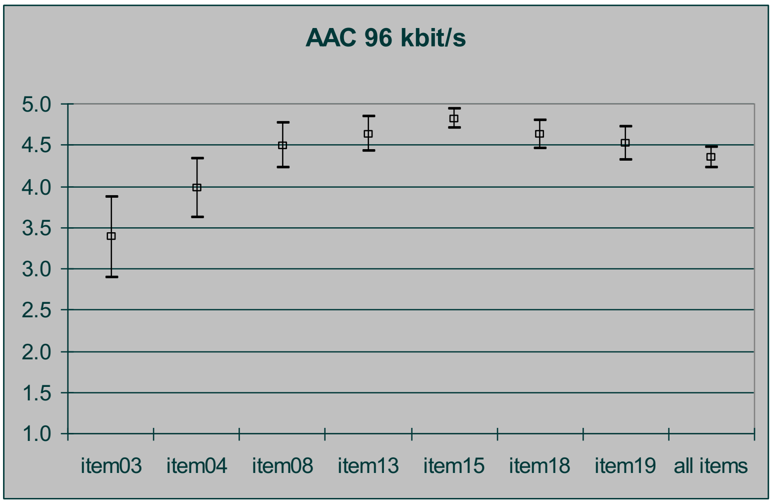


Figure 6-36: Scores for AAC 96 kbit/s

6.2.3 Discussion

6.2.3.1 Overall Results

The following statements are valid based on the mean scores and associated two-sided 95 % confidence intervals:

- AAC main at 64 kbit/s had statistically better performance than both ER BSAC at 64 kbit/s and ER BSAC at 72 kbit/s.
- AAC main at 96 kbit/s was not statistically different from ER BSAC at 96 kbit/s.
- ER BSAC 96 kbit/s had statistically better performance than ER BSAC at 88 kbit/s.
- ER BSAC 88 kbit/s was not statistically different from ER BSAC at 80 kbit/s (although this was by a very small margin).
- ER BSAC 80 kbit/s had statistically better performance than ER BSAC at 72 kbit/s.
- ER BSAC 72 kbit/s had statistically better performance than ER BSAC at 64 kbit/s.

Therefore one can conclude that

- At the high end of the tested set of rates, 96 kbit/s, the bit rate scalable feature of ER BSAC does not require any overhead in bit rate in order to achieve a quality comparable to AAC main at 96 kbit/s.
- For the most part, the performance of ER BSAC for the tested set of rates was monotonic with bit rate (i.e. incrementally higher rate resulted in incrementally higher performance).
- BSAC at 64 kbit/s does not perform as well as AAC main profile at 64 kbit/s, and hence BSAC does require some overhead to achieve scalability at the low end of the tested set of rates. BSAC at 72 kbit/s is nearly comparable to AAC main profile at 64 kbit/s, which suggests that the scalability overhead at the low end of the tested set of rates is approximately 12.5 %. (The comparison “nearly comparable” is based on the observation that, for all items except one, the CI for BSAC at 72 kbit/s overlaps the CI for AAC main profile at 64 kbit/s.)

6.2.3.2 Codec-by-Codec Results

In the following tables, the first column indicates a system (codec at a specified bit rate) and the second column associates a number with that system. The numbers, indicating systems, appear again as column headings over the body of the table. In the body of the table, the numeric entries indicate for how many test items the performance of the system in that row is statistically better than the performance of the system in that column. In this test there were a total of 7 test items.

Codec	No.	1	2
ER BSAC 64 kbit/s	1		0
AAC main 64 kbit/s	2	6	

Table 6-5: Number of items with statistically significant differences

Codec	No.	1	2
ER BSAC 96 kbit/s	1		0
AAC main 96 kbit/s	2	0	

Table 6-6: Number of items with statistically significant differences

6.2.3.3 Comparison with earlier Test Results

In the "Audio On Internet" verification test [n2425], conducted in Summer 1998, an earlier version of the ER BSAC coder was assessed. Due to the fact, that different bit rates have been tested in this test, no direct comparison is possible. Nevertheless, a positive tendency can be seen. The main problem of ER BSAC in the previous test was its strong degradation in sound quality while using its scalable feature. This problem could be overcome, i. e. the current test has shown that while downscaling the bit rate the degradation in sound quality is rather moderate.

6.3 Session A3 – ER AAC LD

6.3.1 Analysis Method

After the subjective listening tests were completed, average scores and 95 % confidential intervals were calculated for selected pooling of the data. Specifically, pooling of data was done as follows:

Result	Pooling of data
For each system (Overall Results)	All listeners for all test items for that system
For each item and each system (Codec-by-Codec Results, Item-by-Item Results)	All listeners for that test item and that system

Table 6-7: Pooling of data

In this table “system” refers to a codec at a specific bit rate. The second pooling of the data is presented twice, first in the “Codec-by-Codec Results” section as one plot for each test item, and then in the “Item-by-Item Results” section as one plots for each system. Data from all listeners were used in the analysis.

6.3.2 Results

6.3.2.1 Test Results: 64 kbit/s

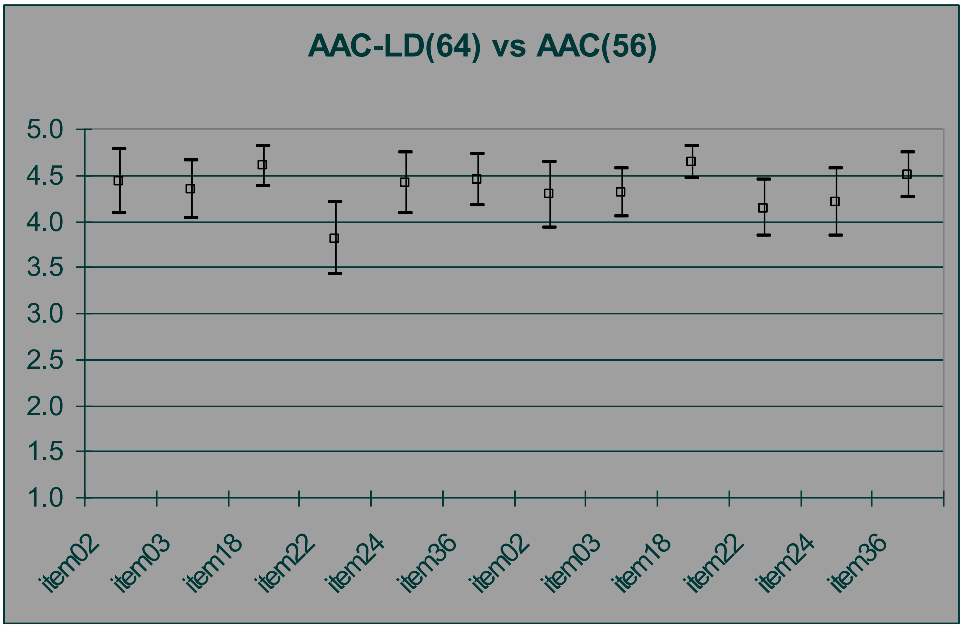


Figure 6-37: Averaged Scores for session A3 – 64 kbit/s
(left 6 scores are for ER AAC LD at 64 kbit/s and right 6 scores are for AAC main at 56 kbit/s)

Average scores for the two systems are as follows:

items	CODEC	mean
ER AAC LD 64 kbit/s	all items	**4.338**
AAC main 56 kbit/s	all items	**4.341**

Table 6-8: Average scores for session A3 – 64 kbit/s

6.3.2.2 Overall Results: 32 kbit/s

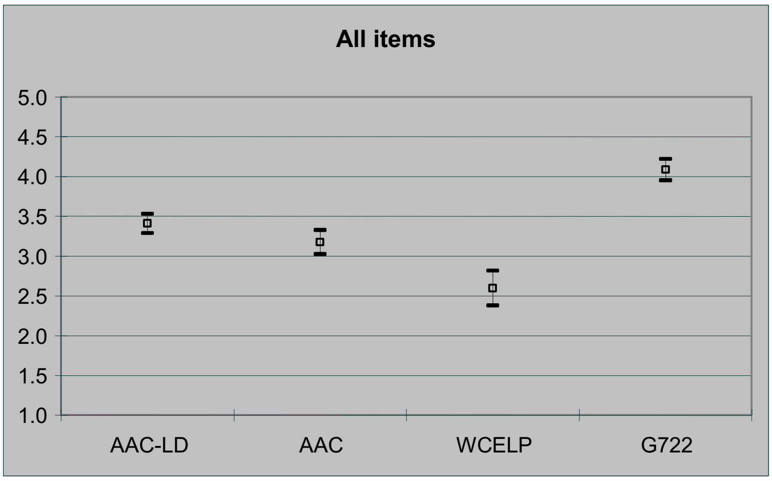


Figure 6-38: Averaged scores for all items

6.3.2.3 Codec-by-Codec Results: 32 kbit/s

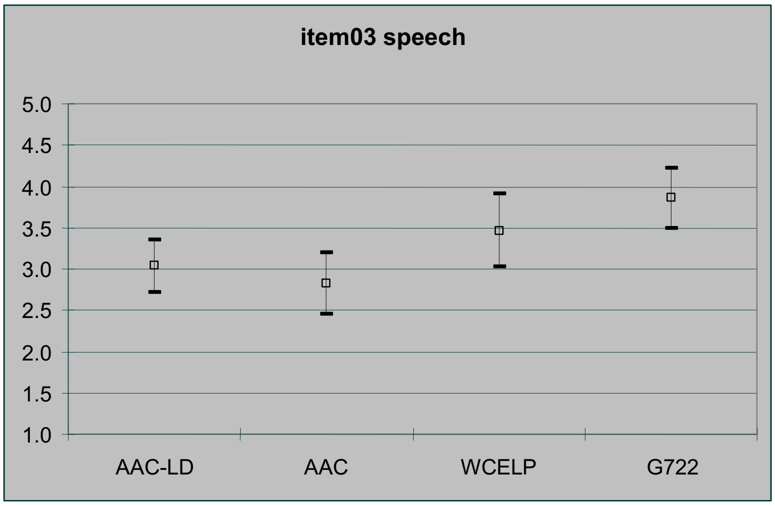


Figure 6-39: Item by item scores (item03)

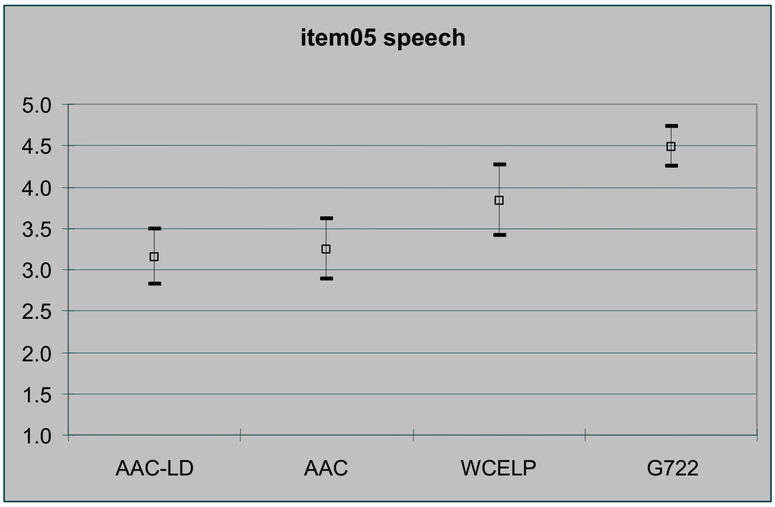


Figure 6-40: Item by item scores (item05)

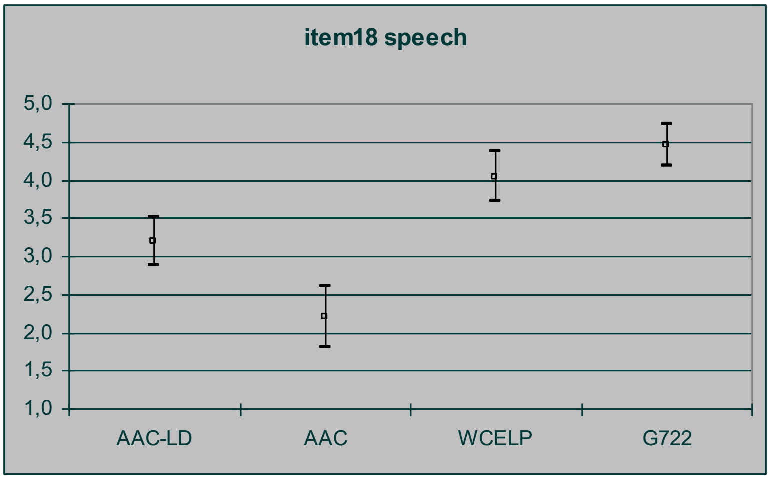


Figure 6-41: Item by item scores (item18)

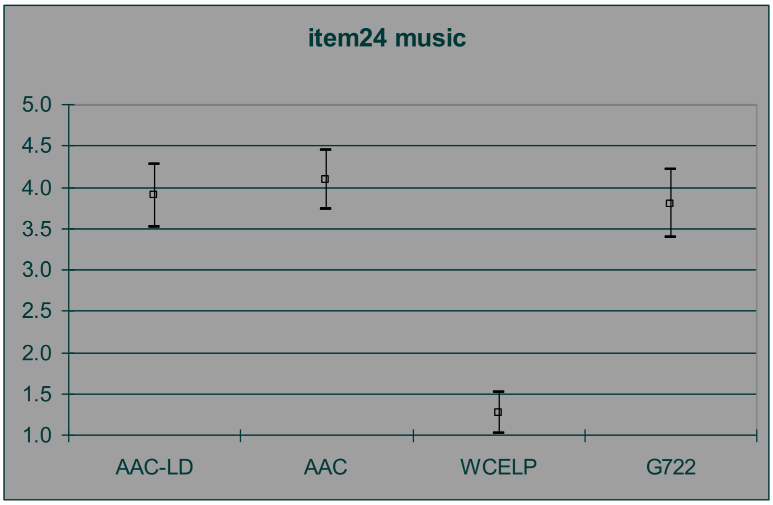


Figure 6-42: Item by item scores (item24)

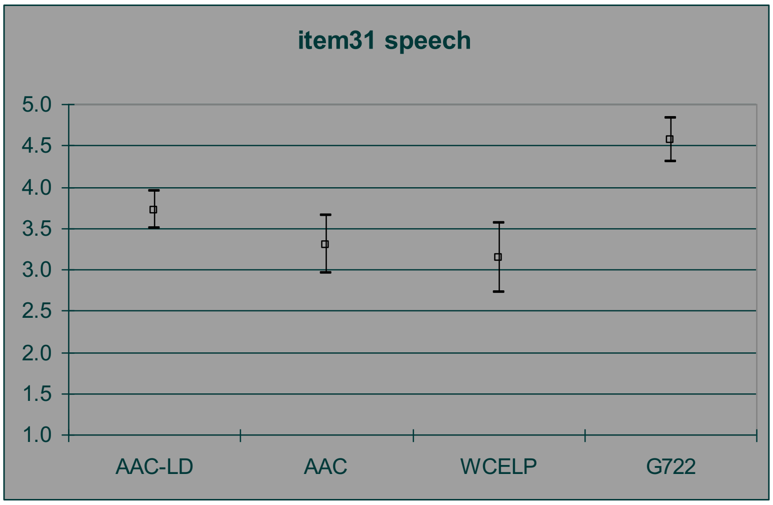


Figure 6-43: Item by item scores (item31)

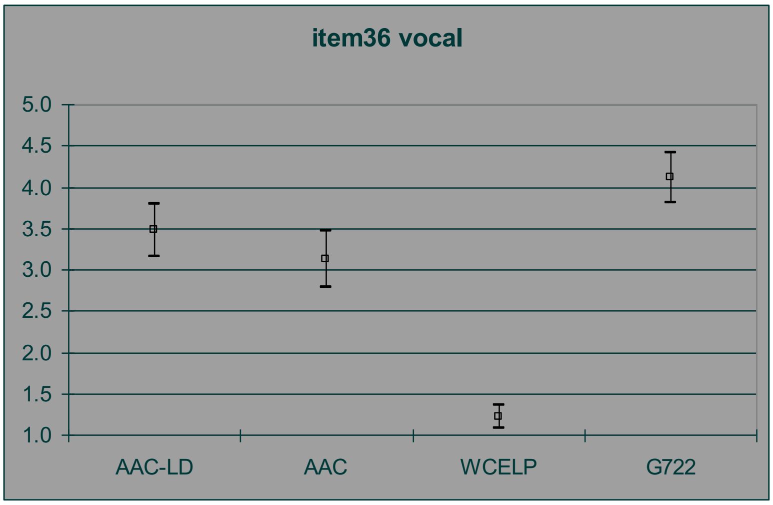


Figure 6-44: Item by item scores (item36)

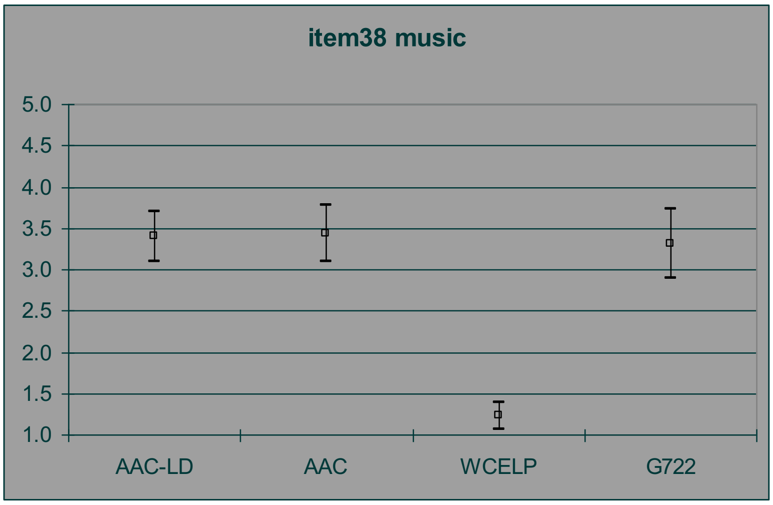


Figure 6-45: Item by item scores (item38)

6.3.2.4 *Item-by-Item Results: 32 kbit/s*

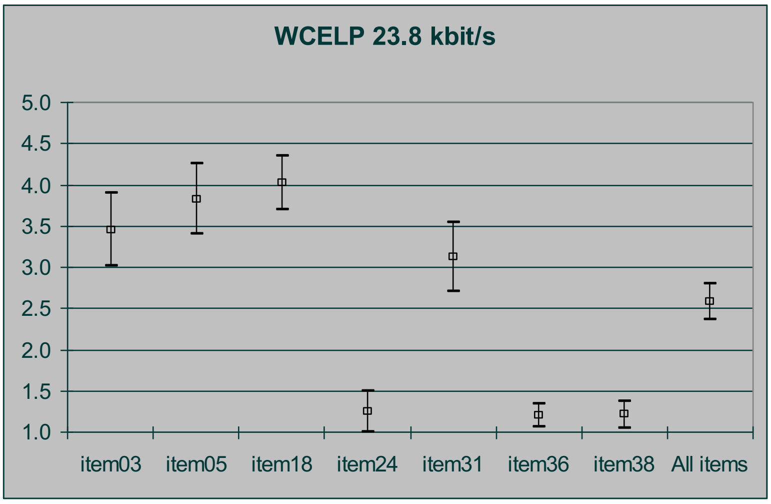


Figure 6-46: Scores for CELP (23.8 kbit/s)

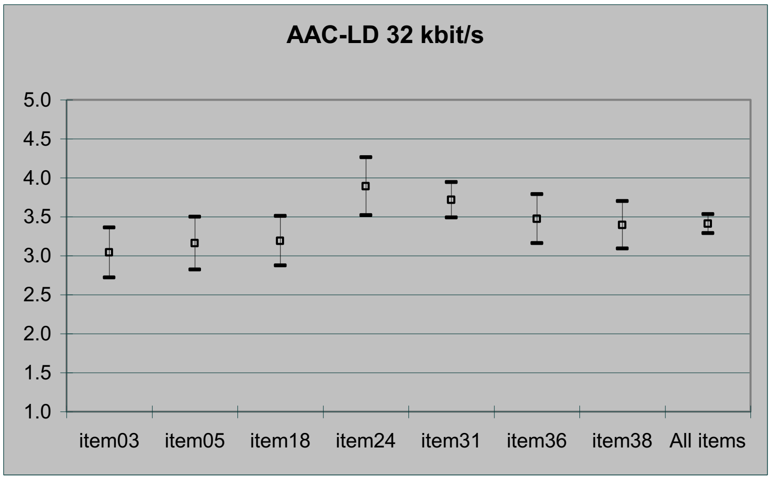


Figure 6-47: Scores for AAC LD (32 kbit/s)

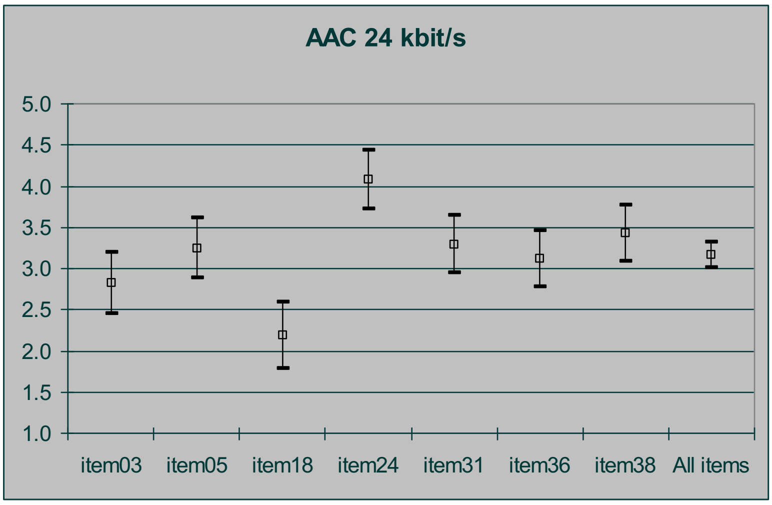


Figure 6-48: Scores for AAC main (24 kbit/s)

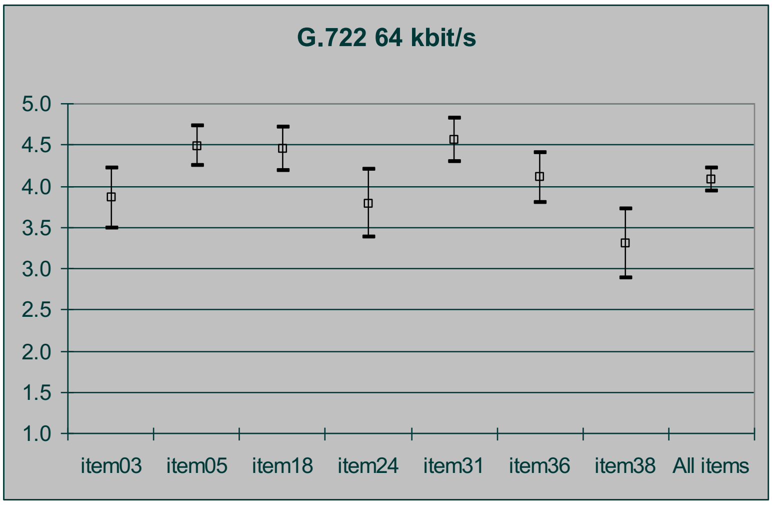


Figure 6-49: Scores for ITU-T G.722 (64 kbit/s)

6.3.3 Discussion

6.3.3.1 Overall Results

The following statements are valid based on the mean scores and associated two-sided 95 % confidence intervals:

- In session A3 @ 64 kbit/s ER AAC LD at 64 kbit/s was not statistically different from AAC main at 56 kbit/s.
- In session A3 @ 32 kbit/s G.722 at 64 kbit/s had statistically better performance than all other systems in this test.
- In session A3 @ 32 kbit/s ER AAC LD at 32 kbit/s was not statistically different from AAC main at 24 kbit/s.
- In session A3 @ 32 kbit/s CELP at 24 kbit/s had statistically worse performance than all other systems in this test.

However, with respect to session A3 @ 32 kbit/s it must be noted that if one considers only the speech items in this test (item03, item05, item18, and item31), the following is true:

- G.722 at 64 kbit/s had statistically better performance than all other systems in this test.
- ER AAC LD at 32 kbit/s was not statistically different from AAC main at 24 kbit/s.
- CELP at 24 kbit/s had statistically better performance than both ER AAC LD at 32 kbit/s and AAC main at 24 kbit/s.

Likewise, if one considers only the music items in this test (item24, item36, and item38), the following is true:

- G.722 at 64 kbit/s had statistically better performance than all other systems in this test.
- ER AAC LD at 32 kbit/s was not statistically different from AAC main at 24 kbit/s.
- CELP at 24 kbit/s had statistically worse performance than all other systems in this test.

The following two graphs analyze the results when separating music and speech items.

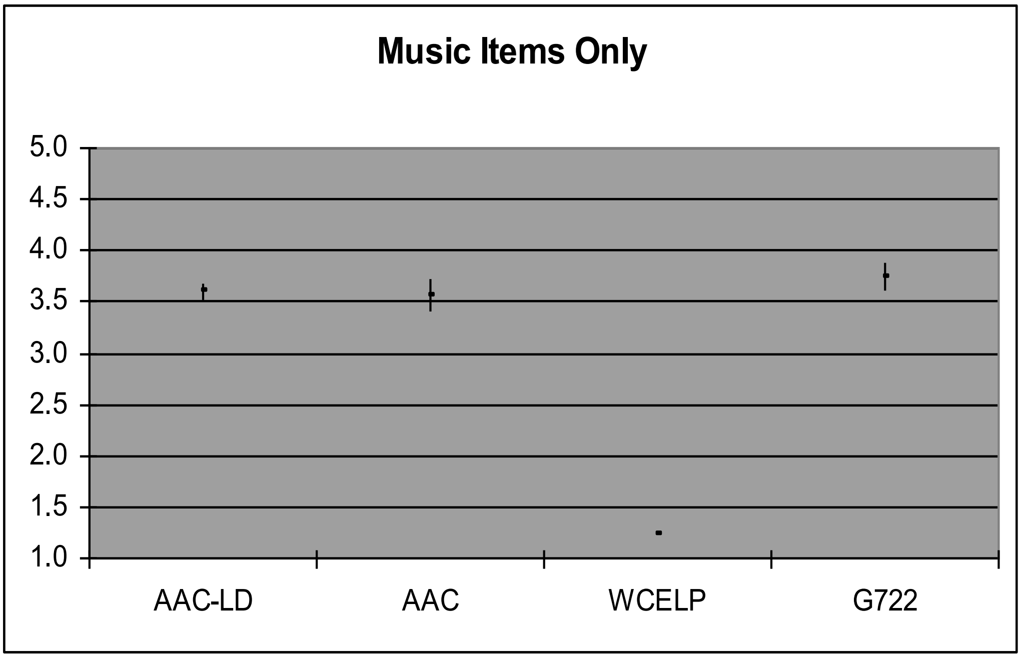


Figure 6-50: Average scores for all music items (session part A3 @ 32 kbit/s)

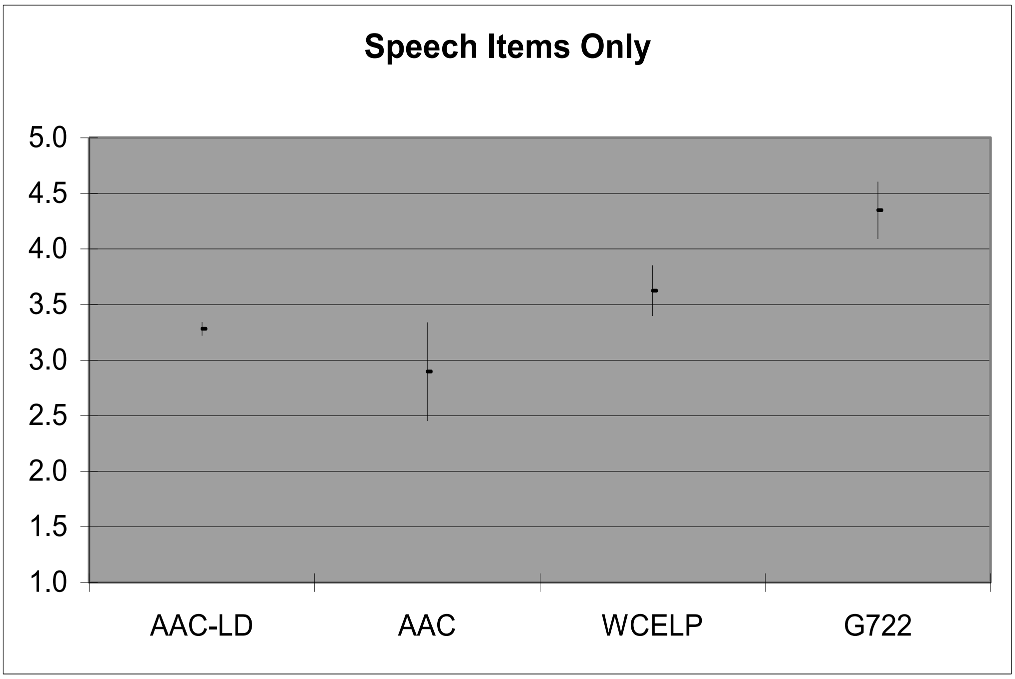


Figure 6-51: Average scores for all speech items (session part A3 @ 32 kbit/s)

Therefore one can conclude that

- ER AAC LD at 64 kbit/s provides performance comparable to that of AAC main at 56 kbit/s, so that a reduction in delay of 86 % (one-way delay reduced from 146 ms for AAC main to 20 ms for ER AAC LD) comes at a cost of an increase in bit rate of approximately 14 % (increased from 56 kbit/s to 64 kbit/s).
- ER AAC LD at 32 kbit/s and 32 kHz sampling rate provides performance comparable to that of AAC main at 24 kbit/s and 24 kHz sampling rate, so that a reduction in delay of 91 % (one-way delay reduced from 323 ms for AAC main to 30 ms for ER AAC LD) comes at a cost of an increase in bit rate of approximately 33 % (increased from 24 kbit/s to 32 kbit/s).
- For unrestricted applications (i.e. for general audio signals, including both music and speech), ER AAC LD provides better performance than CELP.
- However, for applications that are restricted to speech signals only, the CELP coder has a higher performance, a lower delay (15 ms vs. 30 ms) and a lower bit rate (24 kbit/s vs. 32 kbit/s) than the ER AAC LD coder.

6.3.3.2 Codec-by-Codec Results

In the following tables, the first column indicates a system (codec at a specified bit rate) and the second column associates a number with that system. The numbers, indicating systems, appear again as column headings over the body of the table. In the body of the table, the numeric entries indicate for how many test items the performance of the system in that row is statistically better than the performance of the system in that column. In session A3 @ 64 kbit/s there were a total of 6 test items, while in session A3 @ 32 kbit/s there were a total of 7 test items.

Codec	No.	1	2
ER AAC LD 64 kbit/s	1		0
AAC main 56 kbit/s	2	0	

Table 6-9: Session A3 @ 64 kbit/s

Codec	No.	1	2	3	4
ER AAC LD 32 kbit/s	1		0	2	0
AAC main 24 kbit/s	2	0		2	0
CELP 23.8 kbit/s	3	1	1		0
ITU-T G.722 64 kbit/s	4	5	5	4	

Table 6-10: Session A3 @ 32 kbit/s, All Items

6.4 Session A4 – Error Robustness

6.4.1 Analysis Method

After the subjective listening tests were completed, average scores and 95 % confidential intervals were calculated for selected pooling of the data. Specifically, pooling of data was done as follows:

Result	Pooling of data
For each system	All listeners for all test items for that system
For each item and each system	All listeners for that test item and that system

Table 6-11: Pooling of data

In this table "system" refers to a codec at a specific bit rate. Some listener data was not reliable, and was excluded from the analysis (see Annex C.4).

6.4.2 Results

The results of the two parts of session A4, A4 @ 16 kbit/s and A4 @ 96 kbit/s, from each of the two test sites, NTT DoCoMo and FhG, are presented in the following four graphs. The first 8 sections of the graph (labeled I01, I02, I11, I13, I20, I31, I33, and I36 on the horizontal axis) show the scores for each item and each system. The description of each test item is in section 4.2.4.

The first stroke in the graph section is the first system under test, the second the second system under test, and so on up to the last system. The specification of the system is in section 5.2, and is repeated here:

A4 @ 16 kbit/s:	**A4 @ 96 kbit/s:**
1. full bandwidth hidden reference 2. low pass filtered hidden reference (7 kHz) 3. low pass filtered hidden reference (3.5 kHz) 4. low pass filtered hidden reference (1.7 kHz) 5. undistorted (clear channel condition) 6. distorted (critical channel condition) 7. distorted (very critical channel condition)	1. full bandwidth hidden reference 2. low pass filtered hidden reference (7 kHz) 3. low pass filtered hidden reference (3.5 kHz) 4. undistorted (clear channel condition) 5. distorted (critical channel condition) 6. distorted (very critical channel condition)

For both parts of the session A4, A4 @ 16 kbit/s and A4 @ 96 kbit/s, the last section of the graphs shows the overall scores for each system when averaged over all listeners and all test items.

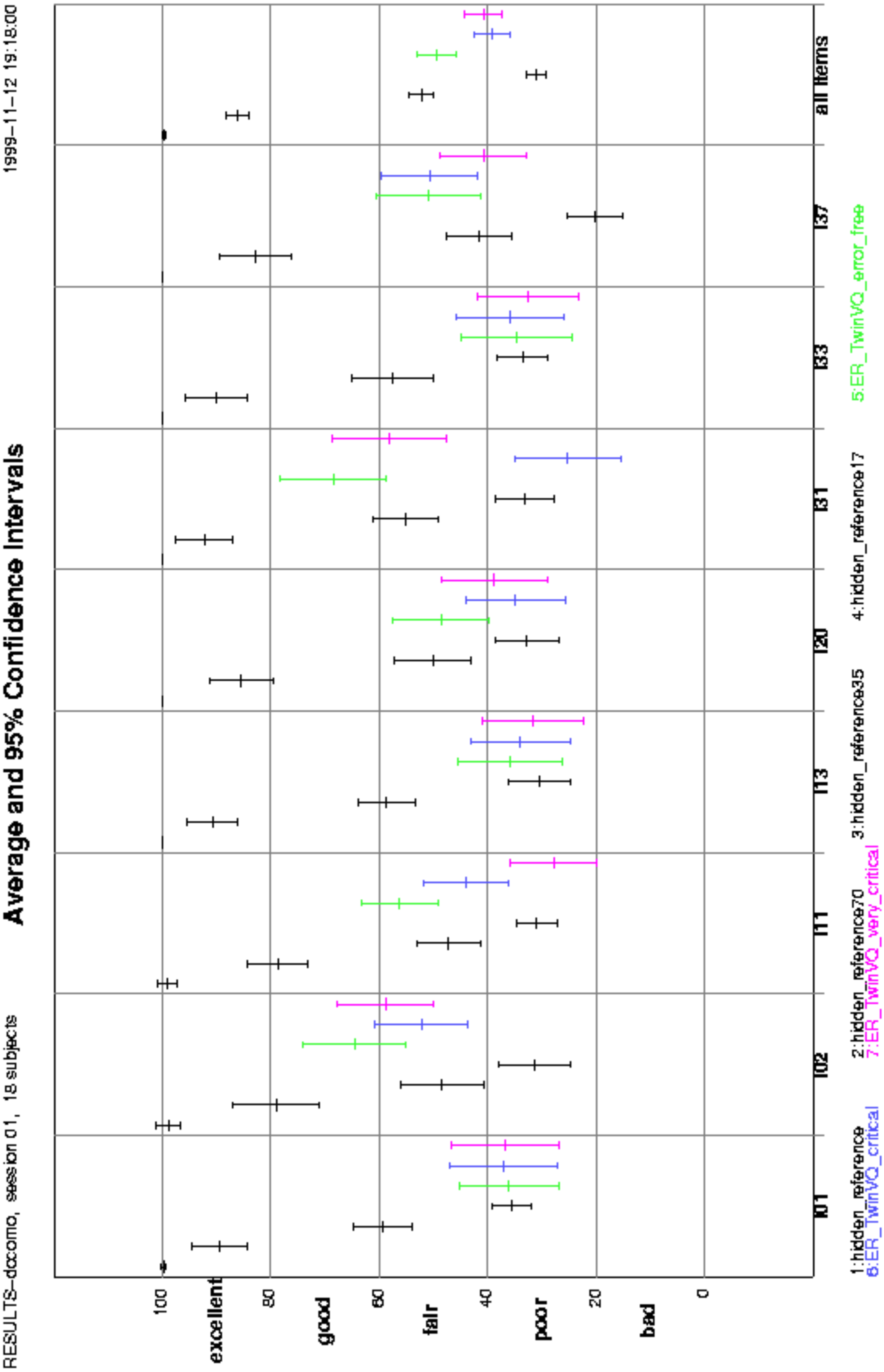


Figure 6-52: Scores for session A4 – 16 kbit/s at NTT DoCoMo

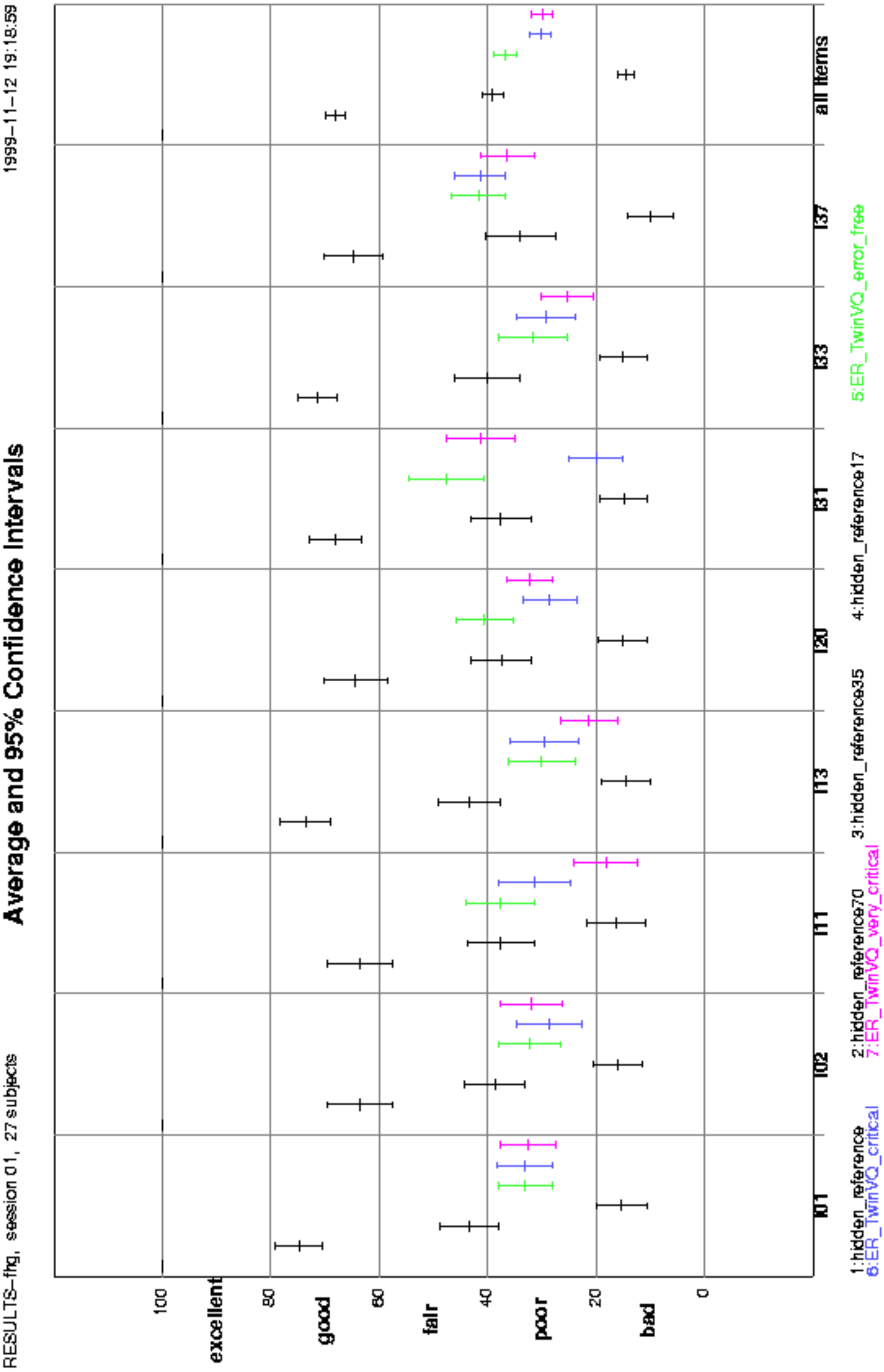


Figure 6-53: Scores for session A4 – 16 kbit/s at FhG

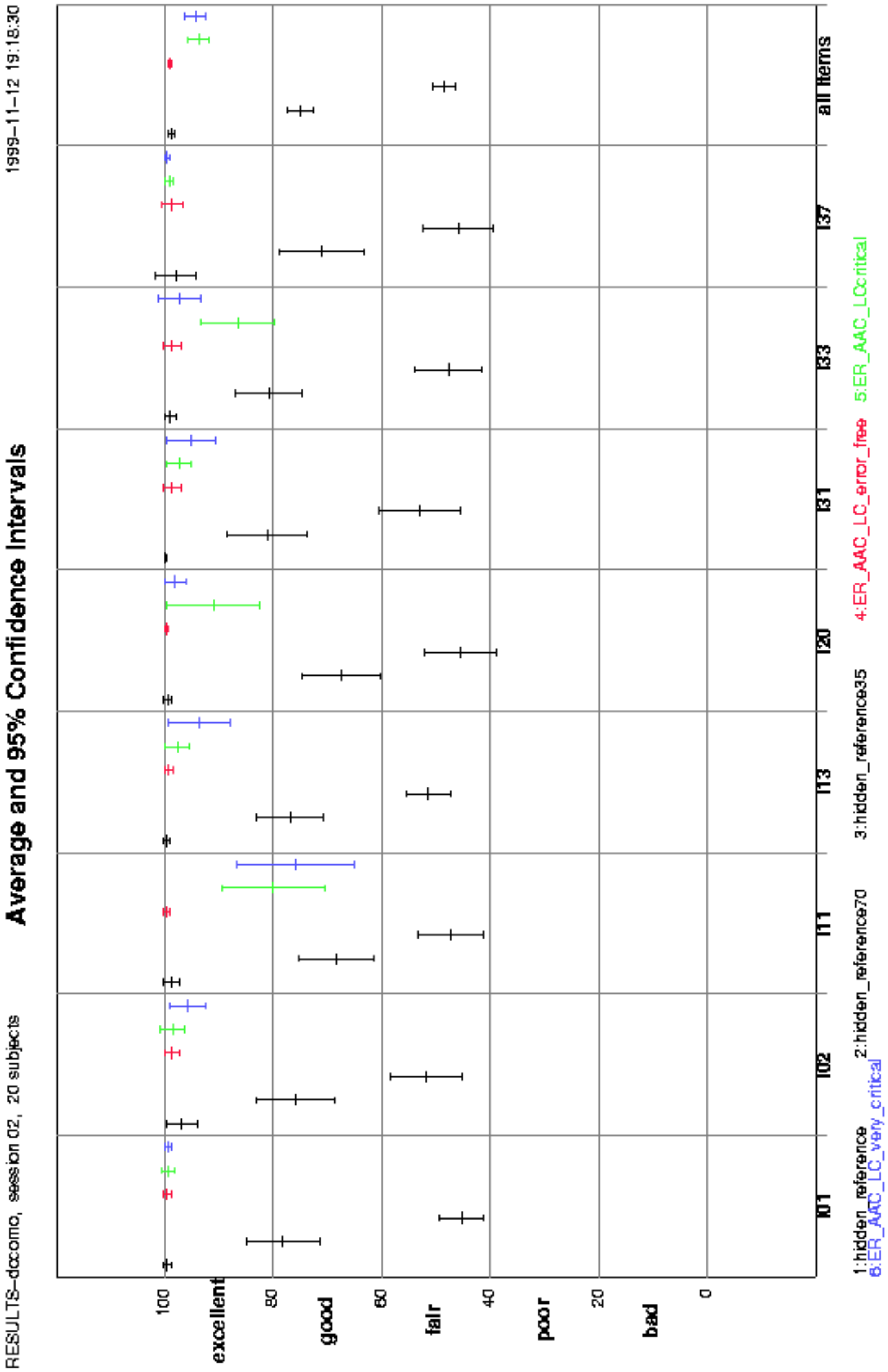


Figure 6-54: Scores for session A4 – 96 kbit/s at NTT DoCoMo

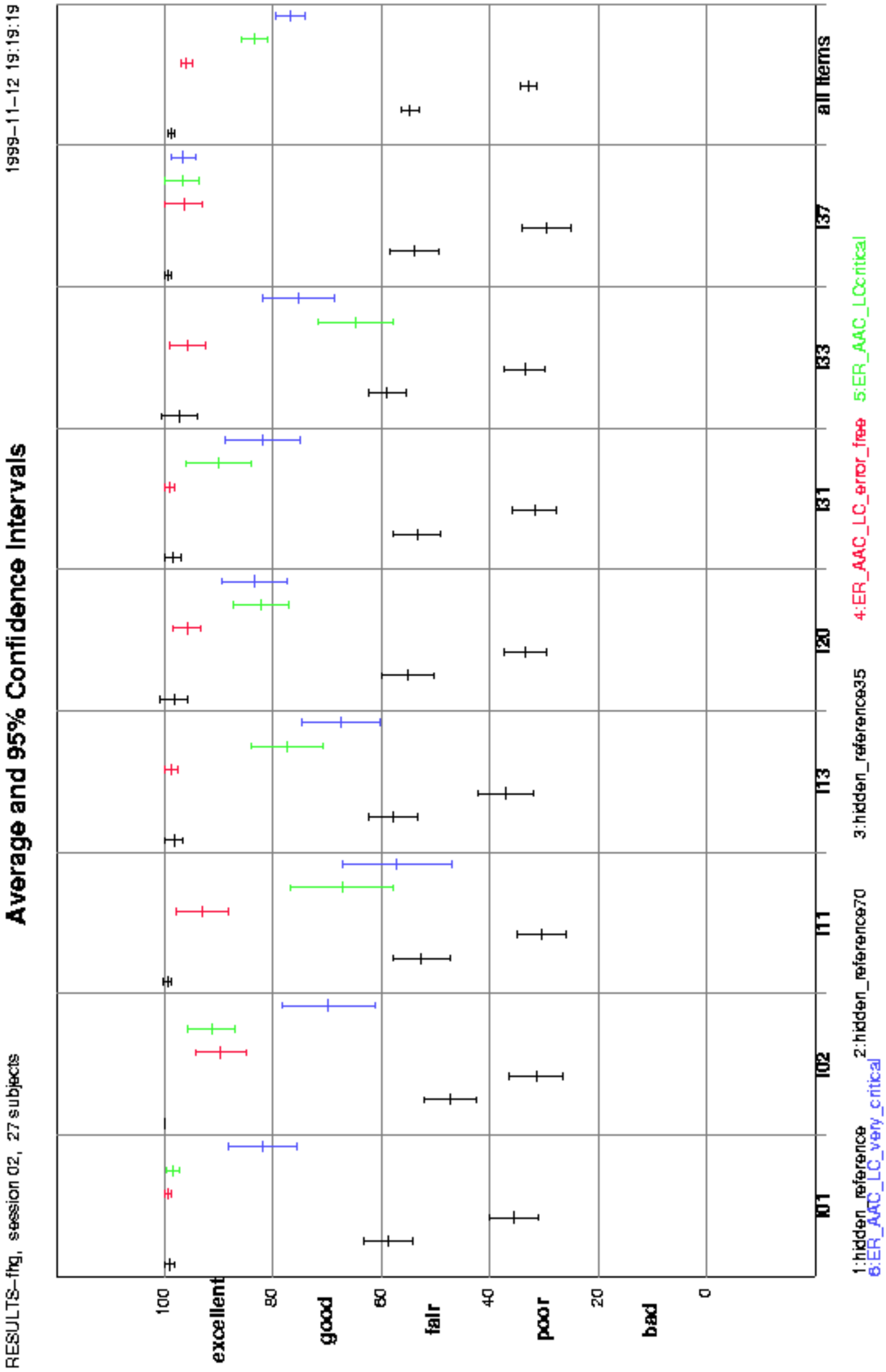


Figure 6-55: Scores for session A4 – 96 kbit/s at FhG

6.4.3 Discussion

6.4.3.1 Overall Results

The following statements are valid based on the mean scores and associated two-sided 95 % confidence intervals (CI):

- In all tests and test sites except A4 @ 96 kbit/s at FhG, the error-prone channel systems were not statistically different from each other.
- In A4 @ 16 kbit/s for both the NTT DoCoMo and FhG test sites, the 95 % CIs of both error-prone channel systems were between the 95 % CI of the 3.5 kHz bandwidth reference and the 1.7 kHz bandwidth reference.
- In A4 @ 96 kbit/s for both the NTT DoCoMo and FhG test sites, the 95 % CIs of both error-prone channel systems were between the 95 % CI of the full bandwidth reference and the 7.0 kHz bandwidth reference.

The following are also valid (but not surprising) statements:

- In both A4 @ 16 kbit/s and A4 @ 96 kbit/s, for both the NTT DoCoMo and FhG test sites, the clear channel system had statistically better performance than both error-prone channel systems.
- In both A4 @ 16 kbit/s and A4 @ 96 kbit/s, for both the NTT DoCoMo and FhG test sites, all reference signals had quality scores that were monotonically decreasing with decreasing bandwidth.

Therefore one can conclude that:

- The ER tools provide equivalently good error robustness over the range of channel error conditions used in the test. Hence it appears that the ER tools may be able to address a wide variety of channel error conditions.
- The ER tools provide error robustness with only a modest overhead in bit rate. For the test of ER AAC LD (session A4 @ 96 kbit/s), the total overhead was 9.5 % (2 % for ER & 7.5 % for EP), and for the test of ER TwinVQ (session A4 @ 16 kbit/s), the total overhead was 17 % (EP only).
- In A4 @ 16 kbit/s, the error-prone channel systems had performance better than the 1.7 kHz bandwidth reference but not as good as the 3.5 the kHz bandwidth reference.
- In A4 @ 96 kbit/s, the error-prone channel systems had performance better than the 7.0 kHz bandwidth reference but not as good as the full bandwidth reference.
- Although no statistical statement can be made on this topic, the results suggest that the ER tools provide performance in error-prone channels that is “nearly as good” as the same system operating over a clear channel. This is an especially significant statement for A4 @ 96 kbit/s, in which the clear channel performance is judged to be “excellent.”

Based on the FhG results in A4 @ 96 kbit/s, one can conclude that

- The “very critical channel condition” is a more difficult channel condition that the “critical channel condition” (i.e. the former received a statistically worse score than the latter).

Unfortunately, strong and statistically robust conclusions cannot be drawn from this test data. If any additional tests of the ER and EP tools are conducted, it is suggested to adjust the system bit rates so that it is possible to make strong statistical statements. An example of such a statement is “The performance of Coder A operating at rate R1 over a clear channel is not statistically different from Coder A operating at rate R2 over error-prone channel E,” where Coder A is a coder with ER and EP tools. In this way one can infer the cost, in bit rate, of providing comparable quality service over an error-prone channel using the ER and EP tools as compared to the clear channel conditions that are already well documented in various MPEG-2 and MPEG-4 verification tests.

6.4.3.2 Codec-by-Codec Results

In the following tables, the first column indicates a system (codec at a specified bit rate) and the second column associates a number with that system. The numbers, indicating systems, appear again as column headings over the body of the table. In the body of the table, the numeric entries indicate for how many test items the performance of the system in that row is statistically better than the performance of the system in that column. In both of these tests there were a total of 8 test items. The results are very similar for both test sites.

Codec	No.	1	2	3	4	5	6	7
full bandwidth hidden reference	1		8	8	8	8	8	8
low pass filtered hidden reference (7 kHz)	2	0		8	8	7	8	8
low pass filtered hidden reference (3.5 kHz)	3	0	0		8	3	4	4
low pass filtered hidden reference (1.7 kHz)	4	0	0	0		0	0	0
undistorted (clear channel condition)	5	0	0	0	5		1	1
distorted (critical channel condition)	6	0	0	0	3	0		1
distorted (very critical channel condition)	7	0	0	0	3	0	1	

Table 6-12: NTT DoCoMo Results, A4 @ 16 kbit/s

Codec	No.	1	2	3	4	5	6	7
full bandwidth hidden reference	1		8	8	8	8	8	8
low pass filtered hidden reference (7 kHz)	2	0		8	8	8	8	8
low pass filtered hidden reference (3.5 kHz)	3	0	0		0	0	0	0
low pass filtered hidden reference (1.7 kHz)	4	0	0	0		0	0	0
undistorted (clear channel condition)	5	0	0	0	8		2	1
distorted (critical channel condition)	6	0	0	0	7	0		1
distorted (very critical channel condition)	7	0	0	0	6	0	1	

Table 6-13: FhG Results, A4 @ 16 kbit/s

Codec	No.	1	2	3	4	5	6
full bandwidth hidden reference	1		8	8	0	2	0
low pass filtered hidden reference (7 kHz)	2	0		8	0	0	0
low pass filtered hidden reference (3.5 kHz)	3	0	0		0	0	0
undistorted (clear channel condition)	4	0	8	8		2	1
distorted (critical channel condition)	5	0	6	8	0		0
distorted (very critical channel condition)	6	0	7	8	0	0	

Table 6-14: NTT DoCoMo Results, A4 @ 96 kbit/s

Codec	No.	1	2	3	4	5	6
full bandwidth hidden reference	1		8	8	1	6	7
low pass filtered hidden reference (7 kHz)	2	0		8	0	0	0
low pass filtered hidden reference (3.5 kHz)	3	0	0		0	0	0
undistorted (clear channel condition)	4	0	8	8		5	7
distorted (critical channel condition)	5	0	7	8	0		2
distorted (very critical channel condition)	6	0	6	8	0	0	

Table 6-15: FhG Results, A4 @ 96 kbit/s

7 Conclusions

The MPEG-4 Audio Version 2 coding tools have undergone a performance verification test for coding of monophonic audio signals in the range of 6 kbit/s to 64 kbit/s and stereophonic audio signals in the range of 64 kbit/s to 96 kbit/s. The coding tools tested were Harmonic and Individual Lines plus Noise (ER HILN) coding, Bit Sliced Arithmetic Coding (ER BSAC), Low Delay Advanced Audio Coding (AAC LD) and the Error Robustness tools comprising Error Resilience (ER) and Error Protection (EP). These tools were tested in four distinct tests, and for each of these tests a description of the systems under test, the method of test material selection, the selected test items, the test methodology and the test results were presented.

The results of these tests support the following broad conclusions:

- The base plus enhancement layers of ER HILN support a bit rate scalable coder that provides at all scalable bit rates quality comparable to that of a fixed-rate ER HILN coder at the same bit rate.
- ER HILN has performance comparable to other MPEG-4 coding technology operating at similar bit rates, but provides the additional capability of independent audio signal speed or pitch change while decoding.
- At the upper end of the bit rate range, ER BSAC provides quality comparable to that of AAC main at the same bit rate, and hence the scalability feature comes at no cost to performance. However at the lower end of the range, the scalability provided by ER BSAC appears to require approximately a 12.5 % bit rate overhead relative to AAC main in order for both to deliver comparable quality.
- In the tests ER BSAC demonstrated scalability in approximately 12 % increments, and, for the most part, each increase in rate provided a statistically significant increase in quality.
- At comparable quality levels, ER AAC LD provides a significant decrease in one-way communications delay relative to AAC main, and does so at only a modest increase in bit rate (around 8 kbit/s).
- The test results indicate that the ER and EP tools are able to provide significant error robustness over a range of channel error conditions, and do so with only a modest bit rate overhead.
- The test results suggest that the ER and EP tools enable MPEG-4 coding tools to provide performance in error-prone channels that is nearly as good as the same coding tools operating over a clear channel, even when the clear channel performance approaches the level of "excellent" on the impairment scale.

8 Glossary

AL-PDU	Access Layer – Protocol Data Unit
AL-SDU	Access Layer – Service Data Unit
EP	Error Protection
ER	Error Resilient
LC	Low Complexity
LD	Low Delay
FEC	Forward Error Correction
HCR	Huffman Codebook Reordering (error resilience tool for AAC spectral data, defined in MPEG-4 Audio Version 2)
LCN	Logical Channel Number
MPE	Multi Pulse Excitation
MUX-PDU	Multiplex Layer Protocol Data Unit
MUX-SDU	Multiplex Service Data Unit
RC UCF	Repeat Count Until Closing Flag
RS	Reed-Solomon (FEC block code)
SL	Sync Layer
STL	Software Tools Library
VCB11	Virtual Codebooks (error resilience tool for AAC section data, defined in MPEG-4 Audio Version 2)

9 References

[bs1116] ITU-R: *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems.* Recommendation ITU-R BS.1116, 1994.

[bs1284] ITU-R: *Methods for the Subjective Assessment Of Sound Quality – General Requirements.* Recommendation ITU-R BS.1284, 1996.

[h223] ITU-T: *Multiplexing protocol for low bit rate multimedia communication.* Recommendation H.223, 1996.

[h223B] ITU-T: *Multiplexing protocol for low bit rate multimedia mobile communication over moderate error-prone channels.* Recommendation H.223 Annex B, 1998.

[m2686] Toshio Miki, Toshiro Kawahara and Sanae Hotani, Documentation and Source Code for Channel Error Pattern Generation, JTC1/SC29/WG11, input paper M2686, 41st MPEG meeting, Fribourg, October 1997.

[m4998] Jürgen Herre ; Ralf Geiger ; Ralph Sperschneider: Encoder Configurations Used for MPEG-4 Version 2 Audio Verification Testing. JTC1/SC29/WG11, input paper M4998, 49th MPEG meeting, Melbourne, October 1999.

[m5012] Schuyler Quackenbush: Report on A3 pre-selection, MPEG-4 V2 verification test. JTC1/SC29/WG11, input paper M5012, 49th MPEG meeting, Melbourne, October 1999.

[m5045] Heiko Purnhagen ; Nikolaus Meine: Information on HILN bitstreams for MPEG-4 V2 verification test. JTC1/SC29/WG11, input paper M5045, 49th MPEG meeting, Melbourne, October 1999.

[m5046] Schuyler Quackenbush ; Masayuki Nishiguchi ; Toshio Miki: AHG Report on MPEG-4 Version 2 Audio Verification Tests. JTC1/SC29/WG11, input paper M5045, 49th MPEG meeting, Melbourne, October 1999.

[m5051] Takehiro Moriya ; Takeshi Mori: Configuration of the ER-TwinVQ Objects Submitted to MPEG-4 Version 2 Audio Verification test. JTC1/SC29/WG11, input paper M5051, 49th MPEG meeting, Melbourne, October 1999.

[m5179] Thomas Sporer ; Ralph Sperschneider: Listening Test Procedure for Intermediate Audio Quality. JTC1/SC29/WG11, input paper M5179, 49th MPEG meeting, Melbourne, October 1999.

[m5180] Ralph Sperschneider: A New Listening Test Method Proposed to Test Error Robustness. JTC1/SC29/WG11, input paper M5179, 49th MPEG meeting, Melbourne, October 1999.

[m5273] Thomas Buchholz ; Miikka Vilermo ; Claus Kupferschmidt ; Wiebke Johannsen: Report on the Selection Process for the MPEG-4 Version 2 Audio Verification Tests, Sessions A1 and A2. JTC1/SC29/WG11, input paper M5273, 49th MPEG meeting, Melbourne, October 1999.

[n2157] Laura Contin ; Martin Dietz ; Jean-Bernard Rault: MPEG-4 Audio verification test specifications - NADIB part. JTC1/SC29/WG11, output paper N2157, 43rd MPEG meeting, Tokyo, March 1998.

[n2276] Catherine Colomes ; Caroline Jacobson ; Eric Scheirer: Report on the MPEG-4 audio NADIB verification tests. JTC1/SC29/WG11, output paper N2276, 44th MPEG meeting, Dublin, July 1998.

[n2278] Sang-Wook Kim: MPEG-4 Audio verification tests specifications - Audio on Internet. JTC1/SC29/WG11, output paper N2276, 44th MPEG meeting, Dublin, July 1998.

[n2425] Eric Scheirer ; Sang-Wook Kim ; Martin Dietz: MPEG-4 Audio verification test results: Audio on Internet. JTC1/SC29/WG11, output paper N2425, 45th MPEG meeting, Atlantic City, October 1998.

[n2795] Audio Subgroup: MPEG-4 version-2 Audio verification test workplan. ISO/IEC JTC1/SC29/WG11, output paper N2795, 48th MPEG meeting, Vancouver, July 1999.

[n2953] Audio Subgroup: Workplan for MPEG-4 Version-2 Audio Verification test. ISO/IEC JTC1/SC29/WG11, output paper N2953, 49th MPEG meeting, Melbourne, October 1999.

[n2992] Requirements Group: MPEG-4 Requirements, version 12 (Melbourne revision). ISO/IEC JTC1/SC29/WG11, output paper N2992, 49th MPEG meeting, Melbourne, October 1999.

[n3058] Audio Subgroup: Text of ISO/IEC 14497-3/FDAM 1. ISO/IEC JTC1/SC29/WG11, output paper N3058, 50th MPEG meeting, Maui, December 1999.

A Testing Schedule

Activity	Timeline	# of weeks	Responsibility
test items on ftp	Jul 19 - Aug 02	2	Uni Hannover
codec integration	Jul 19 - Aug 02	2	ER HILN: Uni Hannover ER BSAC: Samsung ER AAC LD: FhG
coding / decoding	Aug 03 - Sep 20	7	ER HILN: Uni Hannover ER BSAC: Samsung ER AAC LD: FhG TwinVQ: NTT AAC LC: FhG G722: Nokia CELP: Philips, NEC
pre-listening / selection of items	Sep 21 - Oct 04	2	A1: T-Nova, Bosch, Nokia A2: T-Nova, Bosch, Nokia A3: AT&T
49th MPEG meeting	Oct 04 - Oct 11	1	
bit stream / bit rate / decoding verification	Sep 21 - Oct 25	5	ER HILN: Samsung ER BSAC: Uni Hannover ER AAC LD: Uni Hannover TwinVQ: Uni Hannover G722: AT&T/Nokia CELP: Philips
upsampling (ResampAudio) / randomization / tape preparation / grading phase (listening) / analysis (average & CI)	Oct 12 - Nov 08	4	A1: Samsung A2: NTT A3: NTT
draft report	Nov 09 - Nov 22	2	A1: AT&T A2: AT&T A3: AT&T

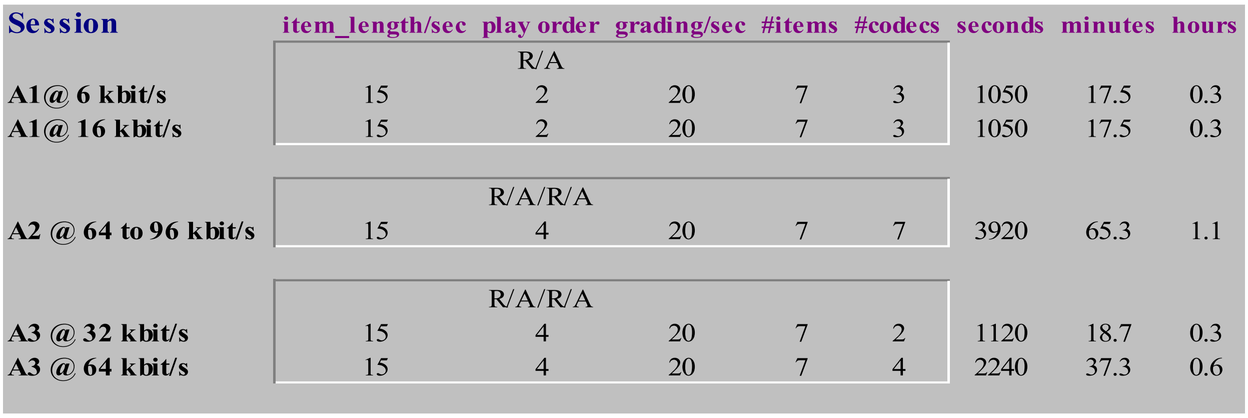
Table A-1: Testing schedule for session A1, A2, and A3

Activity	Timeline	# of weeks	Responsibility
test items on ftp	Jul 19 - Aug 02	2	Uni Hannover
codec integration	Jul 19 - Aug 02	2	ER AAC LC: FhG ER TwinVQ: NTT
coding	Aug 03 - Sep 20	7	ER AAC LC: FhG ER TwinVQ: NTT
channel multiplex / error insertion	Sept 21 - Oct 11	3	NTT DoCoMo
49th MPEG meeting	Oct 04 - Oct 11	1	
decoding	Oct 12 - Oct 18	1	ER AAC LC: FhG ER TwinVQ: NTT
objective measurement / selection of error patterns	Oct 19 - Oct 20	1/2	ER AAC LC: NTT DoCoMo ER TwinVQ: NTT DoCoMo
bit stream / bit rate / decoding verification	Oct 19 - Oct 25	1	ER AAC LC: NTT DoCoMo ER TwinVQ: NTT DoCoMo
upsampling (ResampAudio) / item cutting	Oct 23 - Oct 25	1/2	Uni Hannover
grading phase (listening) / analysis (average & CI)	Oct 26 - Nov 15	3	FhG NTT DoCoMo
draft report	Nov 16 - Nov 22	1	AT&T

Table A-2: Testing schedule for session A4

B Testing Workload

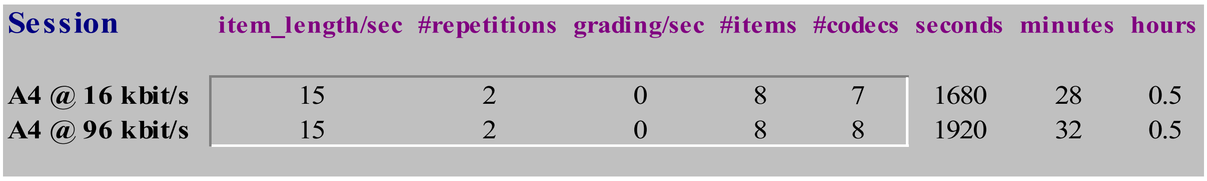
The period of time for listening and grading per listener for session A1, A2, and A3 is listed in the table. Breaks are not included in this calculation.



Session	item_length/sec	play order	grading/sec	#items	#codecs	seconds	minutes	hours
		R/A						
A1@ 6 kbit/s	15	2	20	7	3	1050	17.5	0.3
A1@ 16 kbit/s	15	2	20	7	3	1050	17.5	0.3
		R/A/R/A						
A2 @ 64 to 96 kbit/s	15	4	20	7	7	3920	65.3	1.1
		R/A/R/A						
A3 @ 32 kbit/s	15	4	20	7	2	1120	18.7	0.3
A3 @ 64 kbit/s	15	4	20	7	4	2240	37.3	0.6

Table B-1: Testing workload for session A1, A2, and A3

According to the test method used in session A4 the total listening time per listener depends on his/her time to set the grades (slides) on the display. Therefore the table below is just a rough estimation of the grading period per subject.



Session	item_length/sec	#repetitions	grading/sec	#items	#codecs	seconds	minutes	hours
A4 @ 16 kbit/s	15	2	0	8	7	1680	28	0.5
A4 @ 96 kbit/s	15	2	0	8	8	1920	32	0.5

Table B-2: Testing workload for session A4

C Detailed Information

C.1 Session A1 – ER HILN

Sequence Number	Session A1 : 6 kbps		Session A1 : 16 kbps	
	CODEC	ITEM	CODEC	ITEM
1	TwinVQ 6	12	HILN 16	19
2	HILN0616_BL	20	HILN0616_EL	15
3	HILN0616_BL	38	AAC 16	19
4	HILN 6	39	AAC 16	39
5	TwinVQ 6	7	HILN0616_EL	3
6	HILN0616_BL	39	AAC 16	3
7	HILN 6	38	HILN0616_EL	13
8	HILN 6	7	HILN0616_EL	39
9	TwinVQ 6	38	HILN 16	13
10	TwinVQ 6	29	HILN0616_EL	19
11	HILN0616_BL	12	HILN0616_EL	4
12	HILN0616_BL	11	HILN 16	18
13	HILN 6	20	HILN 16	39
14	HILN0616_BL	29	AAC 16	13
15	TwinVQ 6	39	HILN 16	4
16	TwinVQ 6	20	HILN 16	15
17	HILN 6	29	HILN 16	3
18	HILN0616_BL	7	AAC 16	15
19	HILN 6	12	AAC 16	18
20	TwinVQ 6	11	AAC 16	4
21	HILN 6	11	HILN0616EL	18

Table C-1: Presentation Randomization

CODEC	ITEM	Mean	95 % Confidence Interval	
			Lower Bound	Upper Bound
HILIN 6	Item_07	**1.5625**	1.1922	1.9328
	Item_11	**1.3750**	0.9701	1.7799
	Item_12	**2.6875**	2.2143	3.1607
	Item_20	**2.3125**	1.7907	2.8343
	Item_29	**1.8125**	1.3725	2.2525
	Item_38	**1.6250**	1.2693	1.9807
	Item_39	**1.4375**	1.1484	1.7266
	OVERALL	**1.8304**	1.6659	1.9948
HILN0616_BL	Item_07	**1.7500**	1.3438	2.1562
	Item_11	**1.6875**	1.3838	1.9912
	Item_12	**2.6875**	2.2351	3.1399
	Item_20	**1.8750**	1.4476	2.3024
	Item_29	**1.6250**	1.1565	2.0935
	Item_38	**1.5625**	1.2487	1.8763
	Item_39	**1.2500**	0.9563	1.5437
	OVERALL	**1.7768**	1.6247	1.9289
TwinVQ 6	Item_07	**2.0625**	1.7436	2.3814
	Item_11	**2.1250**	1.6913	2.5587
	Item_12	**2.0625**	1.6635	2.4615
	Item_20	**1.6875**	1.2975	2.0775
	Item_29	**2.1250**	1.6512	2.5988
	Item_38	**1.4375**	1.0779	1.7971
	Item_39	**2.4375**	1.9654	2.9096
	OVERALL	**1.9911**	1.8397	2.1425
HILIN 16	Item_03	**2.5625**	1.9916	3.1334
	Item_04	**2.6250**	2.1606	3.0894
	Item_13	**1.9375**	1.5216	2.3534
	Item_15	**2.9375**	2.4712	3.4038
	Item_18	**3.0625**	2.5603	3.5647
	Item_19	**3.5000**	3.1063	3.8937
	Item_39	**2.3750**	1.8893	2.8607
	OVERALL	**2.7143**	2.5311	2.8975
HILN0616_EL	Item_03	**2.6875**	3.1859	2.1891
	Item_04	**2.5000**	2.9124	2.0876
	Item_13	**1.8750**	2.2639	1.4861
	Item_15	**2.7500**	3.2046	2.2954
	Item_18	**3.0625**	3.6083	2.5167
	Item_19	**3.6250**	4.0902	3.1598
	Item_39	**2.1250**	2.6236	1.6264
	OVERALL	**2.6607**	2.4741	2.8473
AAC 16	Item_03	**2.3125**	1.8091	2.8159
	Item_04	**3.0000**	2.4572	3.5428
	Item_13	**2.5000**	2.1042	2.8958
	Item_15	**2.8125**	2.4013	3.2237
	Item_18	**3.3750**	2.8875	3.8625
	Item_19	**2.9375**	2.5860	3.2890
	Item_39	**3.3750**	2.9448	3.8052
	OVERALL	**2.9018**	2.7331	3.0705

Table C-2: Means and Confidence Intervals

C.2 Session A2 – ER BSAC

items	CODEC	lower	mean	upper
item03	BSAC64	1.158	**1.375**	1.592
	BSAC72	1.054	**1.300**	1.546
	BSAC80	1.182	**1.421**	1.660
	BSAC88	1.679	**1.996**	2.313
	BSAC96	3.636	**3.988**	4.339
	AAC64	1.405	**1.671**	1.937
	AAC96	2.896	**3.383**	3.871
item04	BSAC64	2.926	**3.367**	3.808
	BSAC72	2.937	**3.400**	3.863
	BSAC80	3.578	**3.967**	4.355
	BSAC88	3.890	**4.238**	4.585
	BSAC96	3.965	**4.313**	4.660
	AAC64	2.999	**3.492**	3.984
	AAC96	3.614	**3.975**	4.336
item08	BSAC64	2.701	**3.008**	3.316
	BSAC72	3.330	**3.683**	4.036
	BSAC80	3.933	**4.242**	4.550
	BSAC88	4.063	**4.358**	4.653
	BSAC96	3.979	**4.250**	4.521
	AAC64	3.929	**4.175**	4.421
	AAC96	4.224	**4.496**	4.768
item13	BSAC64	3.006	**3.279**	3.552
	BSAC72	3.860	**4.079**	4.298
	BSAC80	4.417	**4.600**	4.783
	BSAC88	4.332	**4.554**	4.777
	BSAC96	4.456	**4.633**	4.811
	AAC64	4.379	**4.592**	4.804
	AAC96	4.419	**4.633**	4.847
item15	BSAC64	3.064	**3.358**	3.653
	BSAC72	3.433	**3.750**	4.067
	BSAC80	4.063	**4.367**	4.670
	BSAC88	4.327	**4.554**	4.781
	BSAC96	4.589	**4.754**	4.919
	AAC64	3.721	**4.017**	4.312
	AAC96	4.707	**4.821**	4.935
item18	BSAC64	2.513	**2.913**	3.312
	BSAC72	2.768	**3.196**	3.624
	BSAC80	3.489	**3.838**	4.186
	BSAC88	3.945	**4.288**	4.630
	BSAC96	4.288	**4.525**	4.762
	AAC64	3.443	**3.883**	4.324
	AAC96	4.455	**4.625**	4.795
item19	BSAC64	3.021	**3.379**	3.737
	BSAC72	3.639	**3.946**	4.252
	BSAC80	3.895	**4.125**	4.355
	BSAC88	4.231	**4.475**	4.719
	BSAC96	4.219	**4.454**	4.689
	AAC64	4.041	**4.279**	4.517
	AAC96	4.317	**4.521**	4.724
all items	BSAC64	2.797	**2.954**	3.111
	BSAC72	3.155	**3.336**	3.518
	BSAC80	3.609	**3.794**	3.979
	BSAC88	3.900	**4.066**	4.232
	BSAC96	4.316	**4.417**	4.517
	AAC64	3.549	**3.730**	3.910
	AAC96	4.227	**4.351**	4.475

Table C-3: Means and Confidence Intervals

C.3 Session A3 – ER AAC LD

CODEC	items	lower	mean	upper
ER AAC LD 64	item02	4.075	**4.425**	4.775
	item03	4.029	**4.342**	4.655
	item18	4.369	**4.592**	4.814
	item22	3.412	**3.808**	4.205
	item24	4.083	**4.413**	4.742
	item36	4.162	**4.446**	4.729
AAC 56	item02	3.927	**4.283**	4.640
	item03	4.044	**4.304**	4.564
	item18	4.458	**4.633**	4.808
	item22	3.830	**4.133**	4.437
	item24	3.830	**4.196**	4.562
	item36	4.257	**4.496**	4.734

Table C-4: Averaged scores for session A3 @ 64 kbit/s

CODEC	items	lower	mean	upper
ER AAC LD 64	All items		**4.338**	
AAC 56	All items		**4.341**	

Table C-5: Overall Scores for session A3 – 64 kbit/s

items	CODEC	lower	mean	upper
item03	ER AAC LD	2.720	**3.042**	3.363
	AAC	2.456	**2.829**	3.202
	CELP	3.030	**3.471**	3.911
	G722	3.503	**3.867**	4.231
item05	ER AAC LD	2.823	**3.163**	3.502
	AAC	2.890	**3.254**	3.619
	CELP	3.413	**3.842**	4.271
	G722	4.253	**4.492**	4.730
item18	ER AAC LD	2.873	**3.192**	3.510
	AAC	1.798	**2.196**	2.593
	CELP	3.716	**4.038**	4.359
	G722	4.187	**4.454**	4.722
item24	ER AAC LD	3.519	**3.892**	4.265
	AAC	3.730	**4.088**	4.445
	CELP	1.015	**1.263**	1.510
	G722	3.381	**3.796**	4.211
item31	ER AAC LD	3.489	**3.717**	3.944
	AAC	2.946	**3.300**	3.654
	CELP	2.726	**3.142**	3.558
	G722	4.309	**4.571**	4.833
item36	ER AAC LD	3.161	**3.475**	3.789
	AAC	2.787	**3.125**	3.463
	CELP	1.074	**1.217**	1.359
	G722	3.810	**4.113**	4.415
item38	ER AAC LD	3.091	**3.396**	3.700
	AAC	3.100	**3.438**	3.775
	CELP	1.057	**1.225**	1.393
	G722	2.895	**3.313**	3.730
All items	ER AAC LD	3.290	**3.411**	3.532
	AAC	3.023	**3.176**	3.328
	CELP	2.380	**2.599**	2.819
	G722	3.952	**4.086**	4.221

Table C-6: Means and Confidence Intervals for session A3 @ 32 kbit/s

C.4 Session A4 – Error Robustness

C.4.1 Instructions to Listeners in Session A4

The following information should each listener read carefully prior to the listening:

Details with respect to the test methodology:

- test method is MUSHRA (multi stimulus test with hidden reference and anchors)
- using this test several test signals have to be evaluated at the same time
- a slider is available for each test signal, the assessment will be done using these sliders
- the assessment is based on an analog (continuous) scale, any adjustment is valid
- the scale is subdivided into five areas (excellent, good, fair, poor, bad)
- a visible reference is given
- the listener has the possibility to switch between all test signals of the audio signal in any order and as often as it wants
- one of the test signals is the hidden reference, the listener must grade the version that he thinks it is the hidden reference with the maximum quality level
- pressing "register scores" finishes the grading process definitely (the listener should be careful with this button)

Details with respect to the specific test:

- the test consists of two parts: mono and stereo
- in each part eight items (trials) have to be graded (average length of one item is 15 s)
- within the mono part seven test signals (codecs) have to be assessed, while there are six test signals within the stereo part

C.4.2 Post-Screening Phase

In session A4 the following test sets have been removed during the post-screening. Their disqualification is due to one or both of the following:

- Not following the rules of the test
- Hearing sensitivity significantly worse than average.

A4 @ 16 kbit/s	`docomo-m-l01_40`	The listener could not distinguish between the hidden reference and the first anchor and graded both with the maximum value in six from eight conditions.
A4 @ 16 kbit/s	`docomo-f-l07_40`	The listener could not distinguish between the hidden reference and the first anchor and graded both with the maximum value in six from eight conditions.
A4 @ 16 kbit/s	`docomo-f-l12_41`	The listener could not distinguish between the hidden reference and the first anchor and guessed, the hidden reference was graded with the maximum value three times and the first anchor was graded with the maximum value five times.
A4 @ 16 kbit/s	`fhg-n-m-l23_31`	The listener could not distinguish between the first anchor and the second anchor and graded the second anchor four times better than the first anchor. Furthermore the listener had difficulties to detect the third anchor and graded it three times better than the first or the second anchor.
A4 @ 16 kbit/s	`fhg-n-f-l18_34`	The listener did not follow the rule to grade at least one item with the maximum value.*
A4 @ 96 kbit/s	`docomo-m-l03-40`	The listener could not distinguish between the hidden reference and the first anchor and graded both with the maximum value in all eight conditions.
A4 @ 96 kbit/s	`fhg-n-m-l23_31`	The listener could not distinguish between the first anchor and the second anchor and graded the second anchor six times better than the first anchor.
A4 @ 96 kbit/s	`fhg-n-f-l18_34`	The listener did not follow the rule to grade at least one item with the maximum value.*

Note that the numbering of subjects of A4 @ 16 kbit/s and A4 @ 96 kbit/s do not correspond to each other in case of the listening test site NTT DoCoMo.

---

* This error was possible only for test fhg-?-?-l0[1-5]_??. For all other tests the software forced that the listener to grade at least one item with the value 100.

C.4.3 Tables of Means and Confidence Intervals

Items	Codec	lower	mean	upper
item01	hidden_reference	100	**99.8**	100
	hidden_reference70	84	**89.4**	95
	hidden_reference35	54	**59.4**	65
	hidden_reference17	32	**35.5**	39
	ER_TwinVQ_error_free	27	**36.1**	45
	ER_TwinVQ_critical	27	**37.1**	47
	ER_TwinVQ_very_critical	27	**36.8**	47
item02	hidden_reference	97	**98.9**	101
	hidden_reference70	71	**79.1**	87
	hidden_reference35	41	**48.4**	56
	hidden_reference17	25	**31.3**	38
	ER_TwinVQ_error_free	55	**64.6**	74
	ER_TwinVQ_critical	44	**52.2**	61
	ER_TwinVQ_very_critical	50	**58.9**	68
item11	hidden_reference	97	**99.1**	101
	hidden_reference70	73	**78.7**	84
	hidden_reference35	41	**47.3**	53
	hidden_reference17	27	**30.9**	35
	ER_TwinVQ_error_free	49	**56.3**	63
	ER_TwinVQ_critical	36	**44.1**	52
	ER_TwinVQ_very_critical	20	**27.9**	36
item13	hidden_reference	100	**100.0**	100
	hidden_reference70	86	**90.7**	95
	hidden_reference35	53	**58.7**	64
	hidden_reference17	25	**30.5**	36
	ER_TwinVQ_error_free	26	**36.0**	46
	ER_TwinVQ_critical	25	**33.9**	43
	ER_TwinVQ_very_critical	22	**31.7**	41
item20	hidden_reference	100	**100.0**	100
	hidden_reference70	80	**85.4**	91
	hidden_reference35	43	**50.1**	57
	hidden_reference17	27	**32.8**	39
	ER_TwinVQ_error_free	40	**48.7**	58
	ER_TwinVQ_critical	26	**34.8**	44
	ER_TwinVQ_very_critical	29	**38.8**	49
item31	hidden_reference	100	**100.0**	100
	hidden_reference70	87	**92.3**	98
	hidden_reference35	49	**55.1**	61
	hidden_reference17	28	**33.1**	39
	ER_TwinVQ_error_free	59	**68.5**	78
	ER_TwinVQ_critical	15	**25.2**	35
	ER_TwinVQ_very_critical	48	**58.2**	69
item33	hidden_reference	100	**100.0**	100
	hidden_reference70	84	**90.1**	96
	hidden_reference35	50	**57.5**	65
	hidden_reference17	29	**33.5**	38
	ER_TwinVQ_error_free	24	**34.6**	45
	ER_TwinVQ_critical	26	**35.8**	46
	ER_TwinVQ_very_critical	23	**32.6**	42
item37	hidden_reference	100	**100.0**	100
	hidden_reference70	76	**82.8**	89
	hidden_reference35	35	**41.6**	48
	hidden_reference17	15	**20.2**	25
	ER_TwinVQ_error_free	41	**50.9**	61
	ER_TwinVQ_critical	42	**50.7**	60
	ER_TwinVQ_very_critical	33	**40.8**	49
all items	hidden_reference	99	**99.7**	100
	hidden_reference70	83	**86.1**	88
	hidden_reference35	49	**52.3**	54
	hidden_reference17	29	**31.0**	32
	ER_TwinVQ_error_free	45	**49.5**	53
	ER_TwinVQ_critical	35	**39.2**	42
	ER_TwinVQ_very_critical	37	**40.7**	44

Table C-7: Means and Confidence Intervals for session A4 @ 16 kbit/s at NTT DoCoMo

Items	Codec	lower	mean	upper
item01	hidden_reference	100	**100.0**	100
	hidden_reference70	70	**74.9**	79
	hidden_reference35	38	**43.5**	49
	hidden_reference17	11	**15.3**	20
	ER_TwinVQ_error_free	28	**33.1**	38
	ER_TwinVQ_critical	28	**33.2**	38
	ER_TwinVQ_very_critical	27	**32.6**	38
item02	hidden_reference	100	**100.0**	100
	hidden_reference70	57	**63.6**	70
	hidden_reference35	33	**38.7**	44
	hidden_reference17	12	**16.0**	21
	ER_TwinVQ_error_free	27	**32.2**	38
	ER_TwinVQ_critical	23	**28.6**	35
	ER_TwinVQ_very_critical	26	**32.0**	38
item11	hidden_reference	100	**100.0**	100
	hidden_reference70	58	**63.6**	70
	hidden_reference35	31	**37.6**	44
	hidden_reference17	11	**16.3**	22
	ER_TwinVQ_error_free	31	**37.6**	44
	ER_TwinVQ_critical	25	**31.4**	38
	ER_TwinVQ_very_critical	12	**18.2**	24
item13	hidden_reference	100	**100.0**	100
	hidden_reference70	69	**73.6**	78
	hidden_reference35	38	**43.4**	49
	hidden_reference17	10	**14.4**	19
	ER_TwinVQ_error_free	24	**30.1**	36
	ER_TwinVQ_critical	23	**29.5**	36
	ER_TwinVQ_very_critical	16	**21.3**	27
item20	hidden_reference	100	**100.0**	100
	hidden_reference70	58	**64.3**	70
	hidden_reference35	32	**37.4**	43
	hidden_reference17	10	**15.0**	20
	ER_TwinVQ_error_free	35	**40.6**	46
	ER_TwinVQ_critical	24	**28.6**	34
	ER_TwinVQ_very_critical	28	**32.3**	37
item31	hidden_reference	100	**100.0**	100
	hidden_reference70	63	**68.2**	73
	hidden_reference35	32	**37.6**	43
	hidden_reference17	11	**14.9**	19
	ER_TwinVQ_error_free	41	**47.7**	55
	ER_TwinVQ_critical	15	**20.0**	25
	ER_TwinVQ_very_critical	35	**41.3**	48
item33	hidden_reference	100	**100.0**	100
	hidden_reference70	68	**71.5**	75
	hidden_reference35	34	**40.0**	46
	hidden_reference17	11	**15.0**	19
	ER_TwinVQ_error_free	25	**31.7**	38
	ER_TwinVQ_critical	24	**29.2**	35
	ER_TwinVQ_very_critical	20	**25.3**	30
item37	hidden_reference	100	**100.0**	100
	hidden_reference70	59	**64.9**	70
	hidden_reference35	28	**34.0**	40
	hidden_reference17	6	**9.9**	14
	ER_TwinVQ_error_free	37	**41.7**	47
	ER_TwinVQ_critical	37	**41.4**	46
	ER_TwinVQ_very_critical	31	**36.4**	41
all items	hidden_reference	100	**100.0**	100
	hidden_reference70	66	**68.1**	70
	hidden_reference35	37	**39.0**	41
	hidden_reference17	13	**14.6**	16
	ER_TwinVQ_error_free	34	**36.8**	39
	ER_TwinVQ_critical	28	**30.2**	32
	ER_TwinVQ_very_critical	27	**29.9**	32

Table C-8: Means and Confidence Intervals for session A4 @ 16 kbit/s at FhG

Items	Codec	lower	mean	upper
item01	hidden_reference	99	**99.5**	100
	hidden_reference70	71	**78.2**	85
	hidden_reference35	41	**45.3**	49
	ER_AAC_LC_error_free	99	**99.5**	100
	ER_AAC_LCcritical	98	**99.4**	100
	ER_AAC_LC_very_critical	99	**99.3**	100
item02	hidden_reference	94	**96.8**	100
	hidden_reference70	69	**76.0**	83
	hidden_reference35	45	**51.7**	58
	ER_AAC_LC_error_free	97	**98.7**	100
	ER_AAC_LCcritical	96	**98.6**	101
	ER_AAC_LC_very_critical	92	**95.8**	99
item11	hidden_reference	97	**98.8**	100
	hidden_reference70	61	**68.3**	75
	hidden_reference35	41	**47.3**	53
	ER_AAC_LC_error_free	99	**99.6**	100
	ER_AAC_LCcritical	71	**80.0**	90
	ER_AAC_LC_very_critical	65	**75.9**	87
item13	hidden_reference	99	**99.7**	100
	hidden_reference70	71	**77.0**	83
	hidden_reference35	47	**51.4**	55
	ER_AAC_LC_error_free	98	**99.2**	100
	ER_AAC_LCcritical	95	**97.7**	100
	ER_AAC_LC_very_critical	88	**93.7**	99
item20	hidden_reference	99	**99.5**	100
	hidden_reference70	60	**67.4**	75
	hidden_reference35	39	**45.5**	52
	ER_AAC_LC_error_free	99	**99.7**	100
	ER_AAC_LCcritical	83	**91.1**	100
	ER_AAC_LC_very_critical	96	**98.0**	100
item31	hidden_reference	100	**99.9**	100
	hidden_reference70	74	**81.2**	88
	hidden_reference35	46	**53.0**	60
	ER_AAC_LC_error_free	97	**98.7**	100
	ER_AAC_LCcritical	95	**97.4**	100
	ER_AAC_LC_very_critical	91	**95.2**	100
item33	hidden_reference	98	**99.0**	100
	hidden_reference70	75	**80.8**	87
	hidden_reference35	42	**47.8**	54
	ER_AAC_LC_error_free	97	**98.7**	100
	ER_AAC_LCcritical	80	**86.5**	93
	ER_AAC_LC_very_critical	93	**97.2**	101
item37	hidden_reference	94	**98.0**	102
	hidden_reference70	63	**71.2**	79
	hidden_reference35	39	**45.9**	52
	ER_AAC_LC_error_free	97	**98.7**	101
	ER_AAC_LCcritical	99	**99.2**	100
	ER_AAC_LC_very_critical	99	**99.5**	100
all items	hidden_reference	98	**98.9**	99
	hidden_reference70	72	**75.0**	77
	hidden_reference35	46	**48.5**	50
	ER_AAC_LC_error_free	98	**99.1**	99
	ER_AAC_LCcritical	91	**93.7**	95
	ER_AAC_LC_very_critical	92	**94.3**	96

Table C-9: Means and Confidence Intervals for session A4 @ 96 kbit/s at NTT DoCoMo

Items	Codec	lower	mean	upper
item01	hidden_reference	98	**99.0**	100
	hidden_reference70	54	**58.8**	63
	hidden_reference35	31	**35.6**	40
	ER_AAC_LC_error_free	99	**99.4**	100
	ER_AAC_LCcritical	97	**98.4**	100
	ER_AAC_LC_very_critical	76	**81.9**	88
item02	hidden_reference	100	**100.0**	100
	hidden_reference70	42	**47.2**	52
	hidden_reference35	26	**31.5**	37
	ER_AAC_LC_error_free	85	**89.7**	94
	ER_AAC_LCcritical	87	**91.3**	96
	ER_AAC_LC_very_critical	61	**69.8**	78
item11	hidden_reference	99	**99.5**	100
	hidden_reference70	47	**52.6**	58
	hidden_reference35	26	**30.4**	35
	ER_AAC_LC_error_free	88	**93.1**	98
	ER_AAC_LCcritical	58	**67.2**	77
	ER_AAC_LC_very_critical	47	**57.2**	67
item13	hidden_reference	97	**98.3**	100
	hidden_reference70	53	**57.9**	62
	hidden_reference35	32	**37.1**	42
	ER_AAC_LC_error_free	98	**98.7**	100
	ER_AAC_LCcritical	71	**77.5**	84
	ER_AAC_LC_very_critical	60	**67.5**	75
item20	hidden_reference	96	**98.3**	101
	hidden_reference70	50	**55.1**	60
	hidden_reference35	30	**33.5**	37
	ER_AAC_LC_error_free	93	**95.9**	98
	ER_AAC_LCcritical	77	**82.2**	87
	ER_AAC_LC_very_critical	77	**83.4**	89
item31	hidden_reference	97	**98.4**	100
	hidden_reference70	49	**53.4**	58
	hidden_reference35	28	**31.7**	36
	ER_AAC_LC_error_free	98	**99.0**	100
	ER_AAC_LCcritical	84	**90.0**	96
	ER_AAC_LC_very_critical	75	**81.9**	89
item33	hidden_reference	94	**97.3**	101
	hidden_reference70	55	**59.0**	62
	hidden_reference35	30	**33.6**	37
	ER_AAC_LC_error_free	92	**95.7**	99
	ER_AAC_LCcritical	58	**64.8**	72
	ER_AAC_LC_very_critical	69	**75.3**	82
item37	hidden_reference	99	**99.4**	100
	hidden_reference70	49	**53.9**	59
	hidden_reference35	25	**29.6**	34
	ER_AAC_LC_error_free	93	**96.4**	100
	ER_AAC_LCcritical	94	**96.8**	100
	ER_AAC_LC_very_critical	94	**96.6**	99
all items	hidden_reference	98	**98.8**	99
	hidden_reference70	53	**54.7**	56
	hidden_reference35	31	**32.9**	34
	ER_AAC_LC_error_free	94	**96.0**	97
	ER_AAC_LCcritical	80	**83.5**	86
	ER_AAC_LC_very_critical	73	**76.7**	79

Table C-10: Means and Confidence Intervals for session A4 @ 96 kbit/s at FhG