FlexCode

Project N°: FP6-2002-IST-C 020023-0

Project title: *FlexCode*

Instrument: STREP

Thematic Priority: Information Society Technologies

WP5: Final Subjective Quality Tests – report

Due date: 2009-07
Actual submission date: 2009-07

Start date of project: 2006-07-01 Duration: 36 Months

Organisation name of lead contractor for this deliverable: Orange Labs

Revision: 1.9

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Summary

This document contains the detailed results of the final subjective test of the *FlexCode* coder on speech and audio coding matters.

Orange Labs Contributors

Catherine Colomes, Orange Labs – Rennes
Bernard Letertre, Orange Labs – Rennes
Laetitia Gros, Orange Labs – Lannion
Martine Apperry, Orange Labs – Lannion
Carole Rattazzi, Orange Labs – Lannion

Table of contents

I. INTRODUCTION

In the *FlexCode* Project, the work package WP5 is in charge of testing the resulting *FlexCode* coder at different stages of its development.
The first goal of WP5 was to define a methodological approach able to evaluate the overall quality of the *FlexCode* Codec. In addition, experiments had to be conducted for at least two different "versions" of the *FlexCode* codec, a first one produced at mid-term of the project that led to the deliverable 5.1 **[1]**, and the last "versions" at the end of the project, although some informal subjective tests were run in between.
Since these are the final subjective tests, it was important to test different configurations of the *FlexCode* codec.

This deliverable reports the results of the final audio test set of the *FlexCode* codec. It is split in two main parts that reflect the different tested stages of the codec. Each part contains detailed descriptions of the tested codec configurations. In addition, they contain descriptions of the used methodology, the different tested bit-rates and state-of-the-art codecs, the audio excerpts, the statistical analysis and the results.

The procedures concerning the way this final test was conducted are described in **[2]** and **[3]**, in order to reduce the variations among the different involved laboratories.

II. SOURCE CODEC

1. Tested Codecs

In the project, it has been decided to test two versions of the *FlexCode* source coder that are with "Constraint Entropy" using the KLT transform (FCEKLT), and with "Constraint Resolution" using the MDCT transform (FCRMDCT). A full description of those versions can be found in the deliverable 1.1 of the project **[4]**.
Both *FlexCode* codecs will be tested according 2 points of view: Firstly, their intrinsic quality will be evaluated in a **quality test** using the **MUSHRA** methodology **[6]**. Secondly, their channel adaptation part will be tested using a "**continuous evaluation quality**" methodology (ECQ) **[7]**.
In order to set performance of the *FlexCode* coders, it has been decided to test them comparatively to "anchor coders" that are AMR-WB, AMR-WB+, G729.1 and G722.1, accordingly to the bit-rate.
A short description of each of the state-of-the-art coders is following:

AMR-WB:

AMR-WB codec is based on an algebraic code excitation linear prediction (ACELP) technology. This same technology has been utilised in various speech codec standards, such as GSM-EFR (Enhanced Full Rate) (3GPP TS 06.51) and narrowband GSM-AMR (3GPP TS 26.071). Detailed description of the AMR-WB algorithm can be found in the codec specification (3GPP TS 26.171).
The main novelty in AMR-WB is the sub band structure which enables significant savings in complexity and memory consumption. The audio band is split into two frequency bands so that the internal sampling frequency of the core is 12.8 kHz having audio bandwidth of 50-6400 Hz. Separate processing is performed for the frequency range from 6400 to 7000 Hz. The split band structure enables perceptually efficient bit allocation as well as computational advantages: More bits can be allocated to the perceptually important lower band. At low bit rates operation the higher band is synthesised based on the lower band characteristics, while at the highest bit rate additional bits are reserved for coding the high band coding. Another algorithmic advantage of the sub band structure is the fact that with 12.8 kHz sampling, the 20 ms frame contains 256 samples enabling efficient bit level operations and quantisation schemes e.g. for ACELP algorithms.

AMR-WB+:

The chosen version of the Enhanced AMR is the one standardized in the 3GPP group (March 2005). It has proven to give very good results1 at the bit-rates we used (20 kbps). Low frequencies (0-Fs/4 Hz) are encoded/decoded using the "core" encoder/decoder based on switch ACELP codec and TCX codec (Transformed Coded eXcitation). The high frequency signal is encoded with relatively few bits using bandwidth extension method.

[1] AMR-WB+: "a New Audio Coding Standard for 3rd Generation Mobile Audio Services", Makinen, J.; Bessette, B.; Bruhn, S.; Ojala, P.; Salami, R.; Taleb, A.; Acoustics, Speech, and Signal Processing, 2005. (ICASSP '05). IEEE International Conference on Volume 2, March 18-23, 2005 Page(s):1109 - 1112

G729.1: [5]

ITU-T G.729.1 has been standardized by ITU-T in May 2006 to improve voice quality (narrow band voice quality and extension to high wideband voice quality) over widely deployed G.729 based VoIP infrastructures : G.729 codec is one of the most widely deployed VoIP codec especially in Enterprise environment due to high compression efficiency (8 kbps). G.729.1 coding format includes G.729 coding format to inter work with G.729 installed basis at 8 kbps. Purpose is to increase the voice quality up to high quality wideband telephony services over fixed line access with limited impact on existing infrastructure for smooth transition from narrow band to wideband services.
G.729.1 can operate at 12 different bit rates from 32 down to 8 kbps with wideband quality starting at 14kbit/s. This coder is a bit stream interoperable extension of ITU-T G.729 based on three embedded stages: narrowband cascaded CELP coding at 8 and 12 kbps, time-domain bandwidth extension (TDBWE) at 14 kbps, and split-band MDCT coding with spherical vector quantization (VQ) and pre-echo reduction from 16 to 32 kbps. Side information - consisting of signal class, phase, and energy - is transmitted at 12, 14 and 16 kbps to improve the resilience and recovery of the decoder in case of frame erasures. The maximal coder complexity is around 36 WMOPS. Its algorithmic delay goes from 25 to 48.9375 ms depending on coder modes. Contrary to non-embedded coders such as G.722.2 (AMR-WB) or G.722.1, G.729.1 has strong structural constraints (narrowband CELP core coder, embedded bit stream).

ITU-T G.729.1 Recommendation is provided in **[4]**. It includes a detailed description of the codec, a set of test vectors (in G.729.1 Amendment 1 "New Annex A on G.729.1 usage in H.245, plus corrections to the main body and updated test vectors") and the fixed point simulation software in ANSI-C Code. The Low Delay/Low complexity modes are specified in Amendment 3 of Recommendation ITU-T G.729.1 **[5bis]**.
G.729.1 at 32kbps has been specified as additional speech codec to G.722 for DECT new generation. ETSI TR 102 570 states that "G.729.1 is recommended as an optional codec for wideband speech to provide even higher wideband quality and better robustness to packets/frames losses than G.722 at half the bit rate of G.722. This allows a better transport efficiency on the network side and over the DECT air interface (one full slot). In addition, it is seamless interoperable with largely deployed G.729 based VoIP networks and terminals." The optional wideband speech service profile based on G.729.1 at 32 kbps is specified in ETSI TS 102 527-1.
For usage of G.729.1 over IP networks, format of RTP payload is specified in IETF RFC 4749.
The floating point C Code has been standardized in G.729.1 Annex B "New Annex B on a reference floating-point implementation for G.729.1» and published in **[5]**.
A discontinuous transmission system (DTX) with comfort noise generation is specified in G.729.1 Annex C to allow strong reduction of the coding rate during periods with no active speech. The comfort noise generation system generates a silence insertion description each time an update of the ambient background noise parameters is required to maintain the quality of the generated background noise.

G722.1:

ITU-T G.722.1 is a wideband coding algorithm that provides an audio bandwidth of 50 Hz to 7 kHz, operating at a bit rate of 24 kbps or 32 kbps. The coder is based on transform coding, using a Modulated Lapped Transform and operates on frames of 20 ms. Because the transform window length is 40 ms and a 50% overlap between frames, the total algorithmic delay of the coder is 40 ms. Its complexity is around 5 WMOPS.
G.722.1 is specified for hands free operation in systems with low frame loss. Its main application is in audio and video conferencing.

2. Encoding Process

Three bit rates have been chosen to test the *FlexCode* coders: 14kbps, 24kbps and 32kbps. The chosen anchor codecs were set according to those bit-rates. That leaded to run 3 separated tests according the bit-rate.
THE FCR and FCE were tested comparatively to the AMR-WB at 14 and 24kbps (at 24kbps, the G.729.1 codec was tested as well as an anchor codec). As the AMR-WB cannot achieve exactly that two bit-rates, it was set at 14.25kbps and 23.85kbps respectively. The bitrates of the *FlexCode* coders were set as well to 14.25kbps and 23.85kbps respectively. At 32kbps, the *FlexCode* Coders were tested with the AMR-WB+ and G722.1 for the quality test. For the ECQ test, the FCE was tested comparatively to the G.729.1 at 32kbps.
The generation of the encoded sequences meets the requirements of the processing test plan **[3]**.
Details about the test material processing can be found in the report on *FlexCode* test material processing **[9].**

III. "CHANNEL ADAPTATION" TESTS

In principle, the *FlexCode* codec is able to dynamically adapt to channel constraints, which is useful for streaming and transmission purposes. A full description of the channel coder can be found in [7].

The motivation for testing the channel adaptation of the *FlexCode* codec is to test whether bit errors and packet loss lead to a decrease of the overall quality or no, and to check if the *FlexCode* codec can manage channel problems. This test will be split in two parts:

1. *FlexCode* codec FCEKLT at 32kbps in comparison to "state-of-the-art" codec G.729.1 at 32kbps, in a "packets loss" test.
2. *FlexCode* codec FCRMDCT at 32kbps with bit error adaptation in comparison to *FlexCode* codec FCRMDCT at 20kbps for source coding and 12kbps for channel coding without bit error adaptation.

In this dedicated test, the bit errors and packet loss variations are described latter.

1. Test methodology: ECQ

ECQ stands for Continuous Quality Evaluation. The ECQ methodology has recently been standardized at ITU-T Q12/7 (recommendation P.880, May 2004 [7]). It can be used for evaluating the impact of the time fluctuations in the level of artefacts in speech and audio on the instantaneous perceived quality (that is perceived at any instant of an audio sequence) and on the overall perceived quality (at the end of the audio sequence). The method uses a two-part task: first, an instantaneous judgment on a continuous scale with a slider during the audio sequence, and second, an overall judgment on a standard five-category scale at the end of the audio sequence.

Training phase

Prior to the test, subjects undergo training by listening to two sequences. These sequences, 45-sec long, were extracted from the three test items used in the test and cover different quality levels and different quality fluctuations representative of the range of temporal fluctuations and quality levels that the subjects will encounter during the actual test.

User interface

For the continuous judgment, an electronic slider (e.g. variable resistor) connected to a computer is used for recording the instantaneous quality assessment from the subjects. This device has the following characteristics: slider mechanism without any "re-set" position (i.e., no automatic return to a pre-defined position), linear range of travel of 11 cm, fixed on individual test desk. The "slider position" is recorded twice a second (fast enough to accurately capture responses from the subjects), and is coded from 0 (bottom of scale) to 255 (top of scale), which is an acceptable resolution. **The initial slider position was always at the midpoint of the scale**.

For the overall judgement, a set of five buttons, numbered from 1 to 5, is used. These buttons are horizontally positioned and inlaid in the individual task.

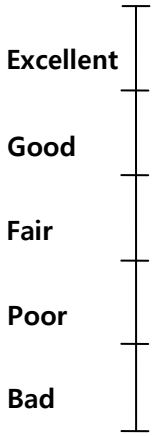
Test instructions and duration

For each sequences, the subjects' task is twofold: a continuous evaluation while listening to the sequence, and an overall evaluation at the end of the sequence.

- Continuous evaluation

Firstly, subjects are instructed to assess the audio quality of the sequence continuously by moving a slider along a continuous scale so that its position reflects their opinion on quality at that instant; the subjects can position the slider anywhere on the scale. Five labels are shown along the scale, i.e.: Excellent, Good, Fair, Poor and Bad to help the subject associate the slider position with suitable ranges of audio quality.

Continuous-quality scale



- Overall evaluation

Secondly, at the end of each sequence, subjects are asked to rate its overall quality on the following 5-category listening-quality scale.

Overall-quality scale

Quality of the speech	*Associated Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

The listening panel

The P.880 recommends at least 24 naïve listeners participating in a test. Listeners were split in 3 groups of 8.

2. Test items and channel variations

Due to the methodology, only three items were chosen: two more speech oriented and another containing music. They were chosen to be realistic types of audio excerpts as much as possible, keeping in mind that they must remain as critical as possible as well. They are listed in table 1:

Name	Category	Description	Time duration	Origin
Nsp_MS2_RefECQ	Speech with noise (basket)	Commentary of a basket ball match with applause and people shouting	1mn10	Orange labs database
Mu_JS1_RefECQ	Music	Jazz music with a female singer in English	1mn30	Orange labs database
Nsp_MS3_RefECQ	Speech with noise (Football)	Commentary of a football play with applause and people shouting	2mn44	Orange labs database

Table 1: Test items

Two different patterns of packets loss variations were applied to those three excerpts (see table 2 and 3), with the *FlexCode* codec.

Packet loss	0%	5%	10%	20%
Mu_JS1_FCEKLT32_pl1 Mu_JS1_G729132_pl1	0 to 25s	25s to 45s	45s to 1mn05	1mn05 to 1mn30
NSp_MS2_FCEKLT32_pl1 NSp_MS2_G729132_pl1	0 to 20s	20s to 35s	35s to 50s	50s to 1mn10
NSp_MS3_FCEKLT32_pl1 NSp_MS3_G729132_pl1	0 to 41s	41s to 1mn22	1mn22 to 2mn03	2mn03 to 2mn44

Table 2: Packet loss variation pattern 1 (profile) for the three audio excerpts

Packet loss	10%	5%	20%	0%
Mu_JS1_FCEKLT32_pl2 Mu_JS1_G729132_pl2	0 to 25s	25s to 45s	45s to 1mn05	1mn05 to 1mn30
NSp_MS2_FCEKLT32_pl2 NSp_MS2_G729132_pl2	0 to 20s	20s to 35s	35s to 50s	50s to 1mn10
NSp_MS3_FCEKLT32_pl2 NSp_MS3_G729132_pl2	0 to 41s	41s to 1mn22	1mn22 to 2mn03	2mn03 to 2mn44

Table 3: Packet loss variation pattern 2 (profile) for the three audio excerpts

The first column indicates the name of the sequences. The value of packet loss is depicted in the first line of each array. Then, the following lines contains the time range during the corresponding value of packet loss is applied on the corresponding sequence, in minutes and seconds, for both tested codecs: FCEKLT (*FlexCode* codec) and G729.1 (State of the art codec).

Two different patterns of bit errors were applied to those three excerpts (see table 4 and 5), with the *FlexCode* codec.

Bit errors	nothing	0dB	4dB	4dB	0dB	nothing
Mu_JS1_FCRMDCT32_be1 Mu_JS1_FCRMDCT32_be1_noadapt	0 to	15s to	30s to	45s to	60s to	1mn15 to 1mn30
NSp_MS2_FCRMDCT32_be1 NSp_MS2_FCRMDCT32_be1_noadapt	0 to	12s to	25s to	35s to	45s to	1mn to 1mn10
NSp_MS3_FCRMDCT32_be1 NSp_MS3_FCRMDCT32_be1_noadapt	0 to	30s to	55s to	1mn20 to	1mn50 to	2mn15 to 2mn44

Table 4: Bit errors variation pattern 1 (profile) for the three audio excerpts

Bit errors	4dB	nothing	0dB	nothing	0dB	4dB
Mu_JS1_FCRMDCT32_be2 Mu_JS1_FCRMDCT32_be2_noadapt	0 to	15s to	30s to	45s to	60s to	1mn15 to 1mn30
NSp_MS2_FCRMDCT32_be2 NSp_MS2_FCRMDCT32_be2_noadapt	0 to	12s to	25s to	35s to	45s to	1mn to 1mn10
NSp_MS3_FCRMDCT32_be2 NSp_MS3_FCRMDCT32_be2_noadapt	0 to	30s to	55s to	1mn20 to	1mn50 to	2mn15 to 2mn44

Table 5: Bit errors variation pattern 2 (profile) for the three audio excerpts

The first column indicates the name of the sequences. The value of bit errors is depicted in the first line of each array. Then, the following lines contain the time range during the corresponding value of bit error is applied on the corresponding sequence, in minutes and seconds, for both tested codecs: FCEKLT (*FlexCode* codec) with channel adaptation, and FCEKLT without channel adaptation.

All in all, 8 conditions were considered (four patterns and 2 codecs per pattern). The first group of naive listeners began to assess the 8 conditions (presented in a random order) with the music sequence, and after a break, they assess the 8 conditions (in different presentation order) with the first speech sequence. After a final break, they assess the 8 last conditions (in different presentation order) with the second speech sequence

The second group tested the first speech sequence first and then the music sequence and finally the second speech item (in a presentation order for the 8 conditions different in the three sessions and different from the presentation order of the first group).

The third group tested the second speech sequence first and then the music sequence and finally the first speech item (in a presentation order for the 8 conditions different in the three sessions and different from the presentation order of the two first groups).

3. Statistical analysis

General analysis

For each subject, if T-s corresponds to the duration of each sequence (in seconds), a data file of 2xT-s values is recorded (i.e., one instantaneous score every 500 ms during T seconds), plus one scalar value (i.e., the overall quality judgment). The 2xT-s instantaneous values (from t=0 until t=2T-1) are subsequently linearly transformed into values from 1 to 5, using the relation S(t) = 1+4 * (slider position/maximum slider position), where S(t) is the instantaneous opinion score. For each sequence, a mean instantaneous judgment (and its standard deviation) is obtained by averaging individual instantaneous judgments over the subjects, at each instant t (i.e., every 500 ms). For each sequence, a mean overall judgment MOS (and its standard deviation) is obtained by averaging individual overall judgments over the subjects on the ACR scale. Substantial deviations between the continuous score and the overall score are likely indications that transient impairments were experienced. This will depend on the recency[2], number of occurrences and duration of the impairments relative to the overall score judgement. Further study is required before this information can be used in transmission planning. Statistical analysis (e.g. ANOVA) can be performed to identify significant effects present in the different experimental conditions.

Post-screening of subjects

According to the recommendation P.880, the responses from subjects should be discarded if those responses exhibit high variations; i.e. if for more than 10% of the time the responses are outside two inter-subject standard deviations (calculated over all subjects), all conditions considered.

4. Results

Overall judgments:

Figures 1 below shows the Mean Opinion Scores and the associated 95-% confidence intervals of the standard deviation, averaged across the 24 subjects, obtained for the speech and the music sequences, respectively.

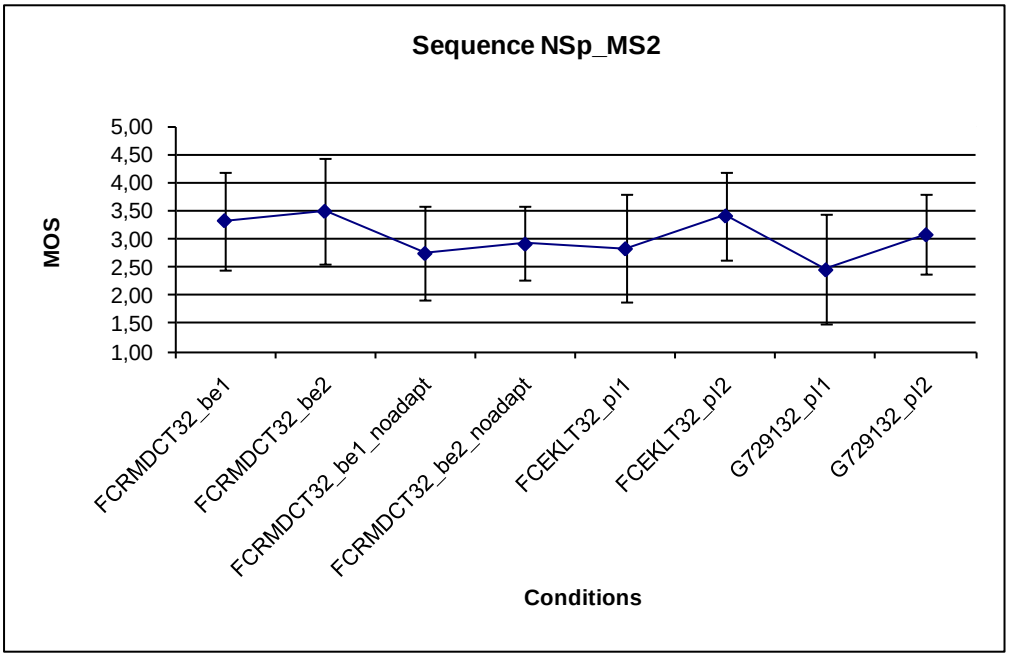


Figure 1: Mean Opinion Scores and associated confidence intervals for the 8 conditions, obtained for the first speech sequence (basket).

[2] Recency effect: Given a list of items to remember, we will tend to remember the last few things more than those things in the middle.

For this speech sequence and bit errors test, the *FlexCode* codec FCMDCT with channel adaptation has been judged to be significantly better (>0.6 on the quality scale) than the *FlexCode* codec FCMDCT without channel adaptation, which is what was expected. This shows that the channel adaptation part is properly working. The same applies for the two other sequences (figures 2 and 3), even more noticeable on the music item.

Comparing the coders in presence of packet loss, the *FlexCode* codec FCEKLT seems to behave a bit better than the G729.1 for both patterns. The difference is 0.4 on the quality scale. This is true as well for the other speech item NSp_MS3 (figure 2) although the difference is lower (around 0.2), and for the music excerpt Mu_JS1 with a difference around 0.3 in favour of the FCEKLT coder.

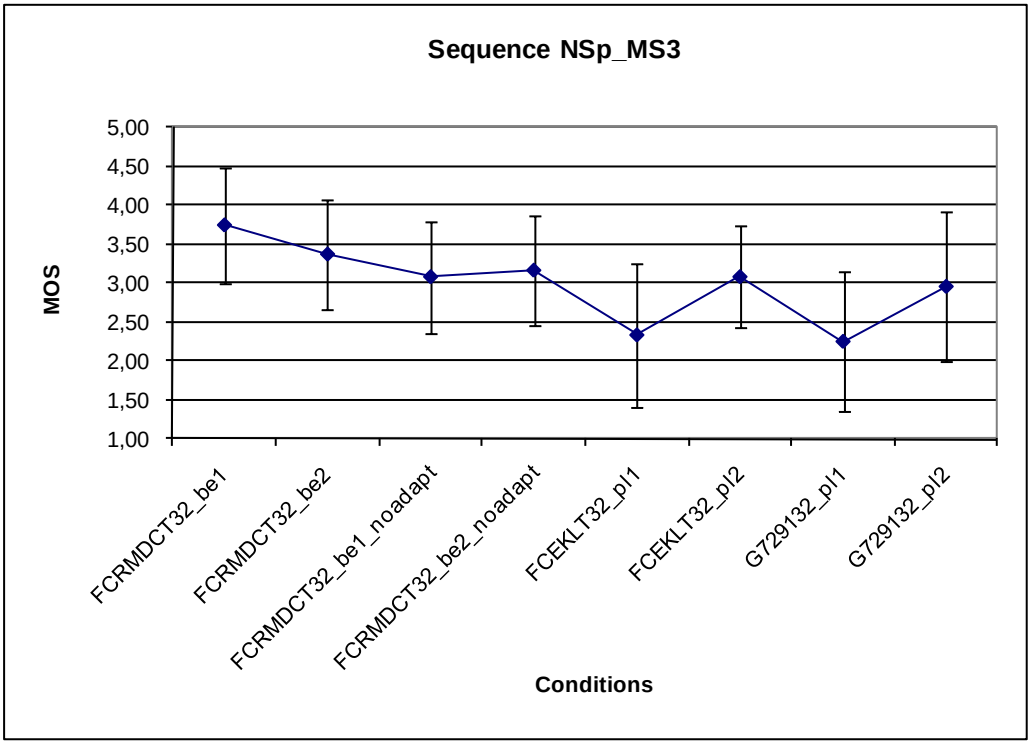


Figure 2: Mean Opinion Scores and associated confidence intervals for the 8 conditions, obtained for the second speech sequence (football).

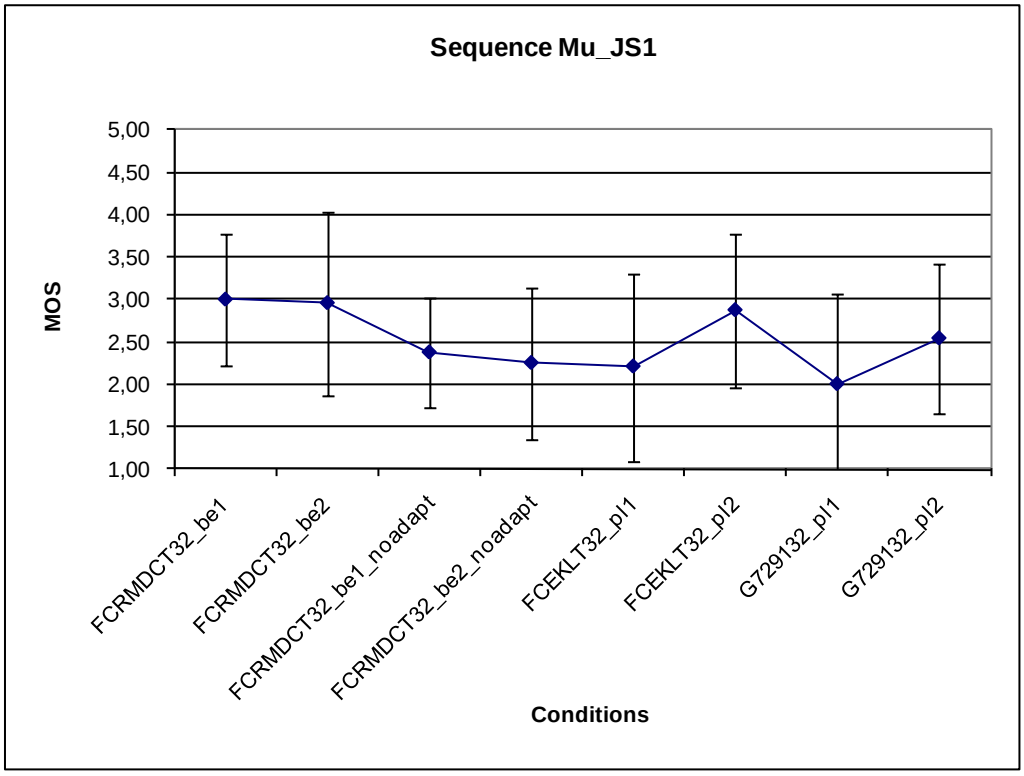


Figure 3: Mean Opinion Scores and associated confidence intervals for the 8 conditions, obtained for the music sequence (jazz).

All those results have to be taken carefully as there is an important overlap of all confidence intervals. They give at least the tendency.

Instantaneous judgments for the "packet loss" conditions

Figure 4 below shows the mean instantaneous responses (the associated standard deviations, which are around 0.5 MOS, are not depicted on the figure for more clarity) for the 2 packet loss conditions (for FCEKLT32 and G729.1@32), obtained with the speech sequence NSp_MS2 (basket).

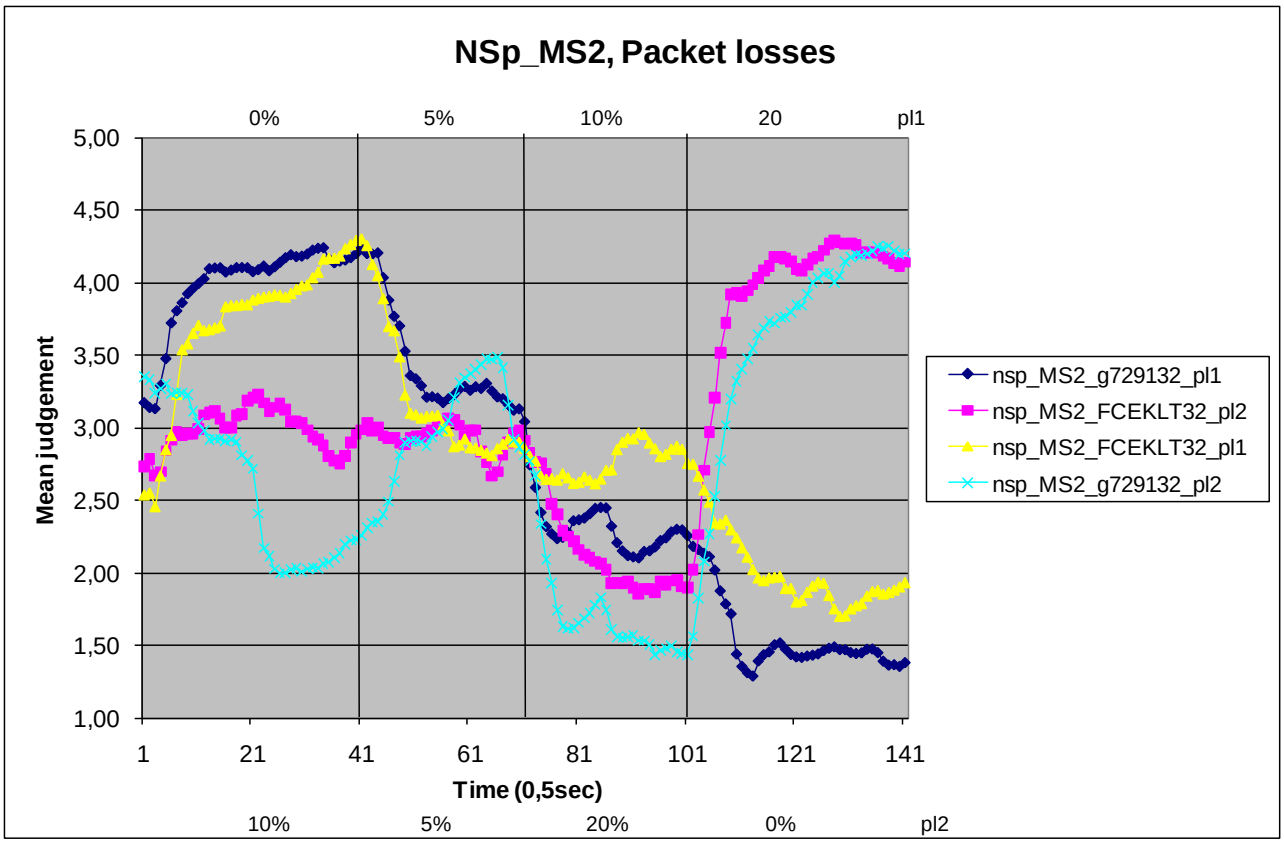


Figure 4: Mean instantaneous responses obtained for 2 packet loss conditions, obtained with the speech sequence NSp_MS2.

Note that all the traces in the figures (4 to 6) start at the same value, because of starting position of the slider (see section "user interface" above).

For pattern 1 (dark and yellow lines), as expected, the quality of both codecs decreases as the packet loss rate increases (from 0% to 20%). We can note that the G729.1 codec performs a bit better in quality than the FCEKLT (0.2 point on the quality scale on the average) for the rates 0 and 5%. Then (for 10 and 20% packet loss rate), it goes the other way as the FCEKLT behaves better than the G729.1 with average scores of 0.5 point above these of the G729.1 on the quality scale.

For pattern 2 (pink and light blue line), the same remarks can be stated : for the 10 and 20% rates, the FCEKLT behaves better than the G729.1 with average scores of 1 point above these of the G729.1 on the quality scale. For the 0 and 5% rates, the difference is lesser, both codecs tending to behave the same.

Remark: There is no difference in quality between 5 and 10% packet loss rate for the FCEKLT (pink line: the first two columns, yellow line: the second and the third column) which shows a good robustness of the *FlexCode* codec at those packet loss rates.

Figure 5 below shows the mean instantaneous responses (the associated standard deviations, which are around 0.5 MOS, are not depicted on the figure for more clarity) for the 2 packet loss conditions (for FCEKLT32 and G729.1@32), obtained with the speech sequence NSp_MS3 (football).

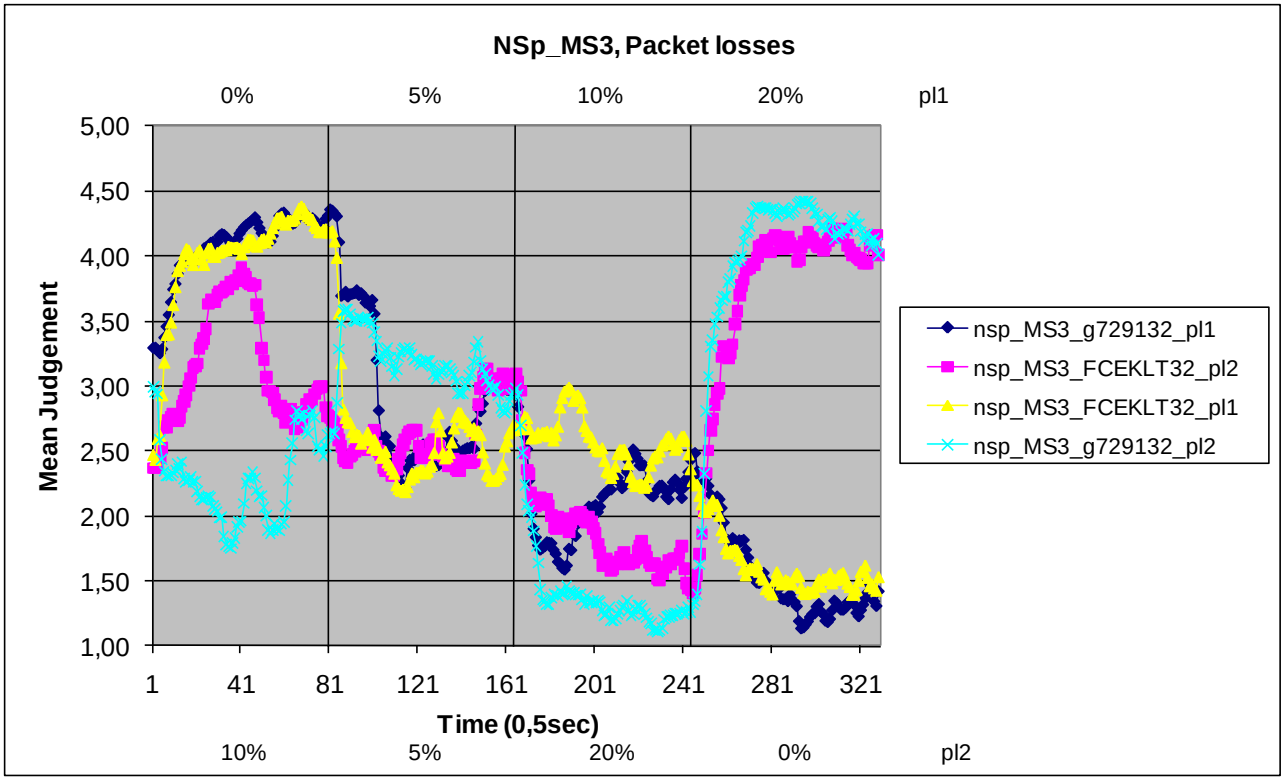


Figure 5: Mean instantaneous responses obtained for 2 packet loss conditions, obtained with the speech sequence NSp_MS3.

For pattern 1 (dark and yellow lines), as expected, the quality of both codecs decreases as the packet loss rate increases (from 0% to 20%). We can note that the G729.1 codec performs the same in quality as the FCEKLT. Note that this is not true around 10% packet loss as the FCEKLT shows the same quality at 5 and 10% (see remark above).

For pattern 2 (pink and light blue line), the same remarks can be stated : for the 10 and 20% rates, the FCEKLT behaves better than the G729.1 with average scores of 0.5 point above these of the G729.1 on the quality scale. For the 0 and 5% rates, the difference is lesser, both codecs tending to behave the same. Some problems of stability have to be noticed at the beginning of the test that put the FCEKLT quality much higher than expected during the first 20s.

Figure 6 below shows the mean instantaneous responses (the associated standard deviations, which are around 0.5 MOS, are not depicted on the figure for more clarity) for the 2 packet loss conditions (for FCEKLT32 and G729.1@32), obtained with the music sequence Mu_JS1 (Jazz).

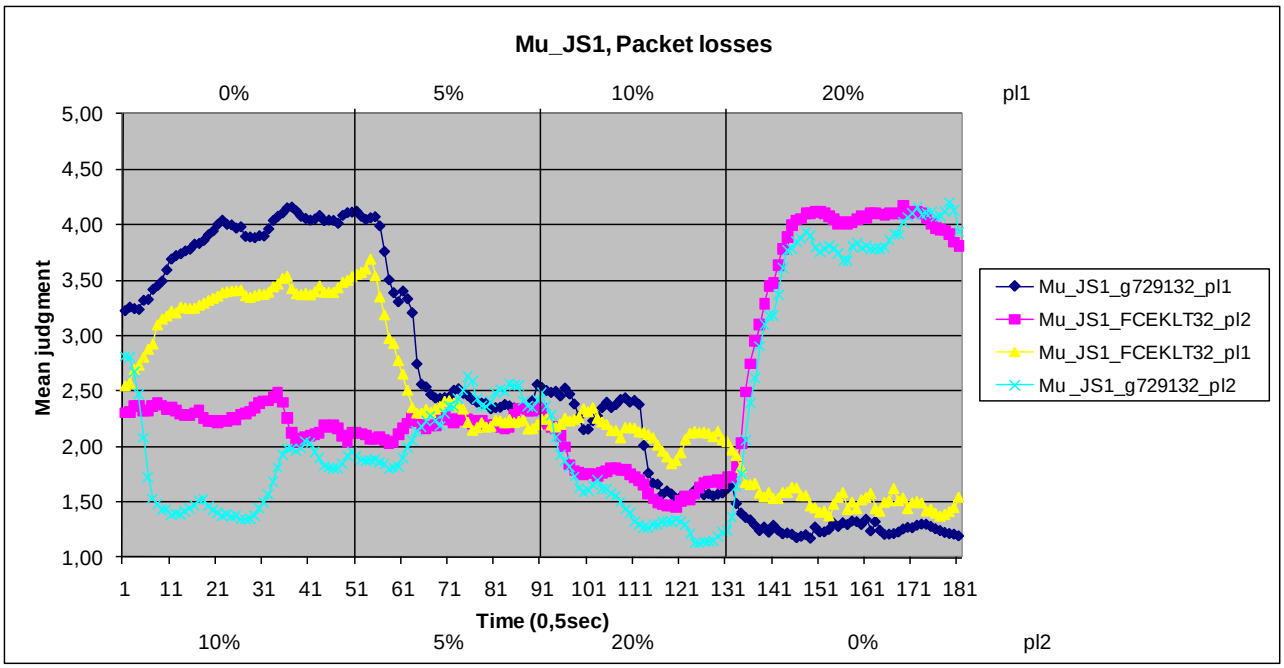


Figure 6: Mean instantaneous responses obtained for 2 packet loss conditions, obtained with the music sequence Mu_JS1.

Basically, the same remarks as those for the speech items can be made for this music items.

Instantaneous judgements for the "bit error" conditions

Figure 7 below shows the mean instantaneous responses (the associated standard deviations, which are around 0.5 MOS, are not depicted on the figure for more clarity) for the 2 bit error conditions (for FCRMDCT with and without channel adaptation), obtained with the speech sequence NSp_MS2 (basket).

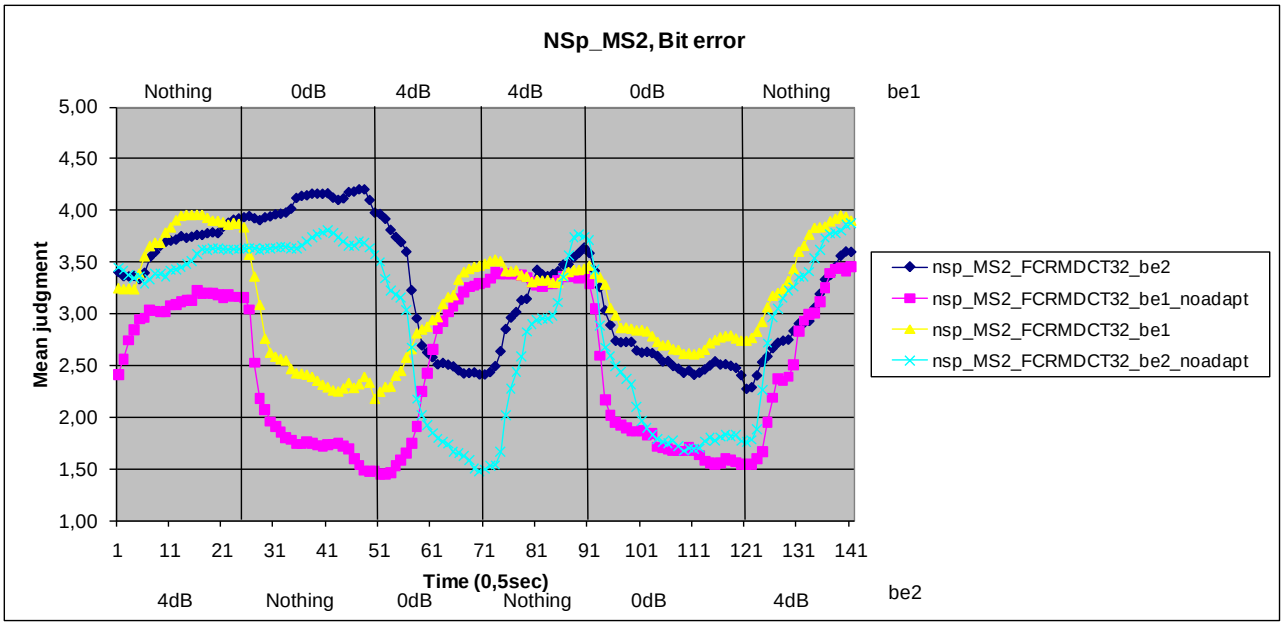


Figure 7: Mean instantaneous responses obtained for 2 packet loss conditions, obtained with the speech sequence NSp_MS2.

Note that all the traces in the figures (7 to 9) start at the same value, because of starting position of the slider (see section "user interface" above).

For pattern 1 (pink and yellow lines) as well as for pattern2 (dark and light blue line), the quality of the FCRMDCT with the channel adaptation has been scored higher than that without channel adaptation, which was expected. The same applied for the two other sequences (figure 8 and 9).

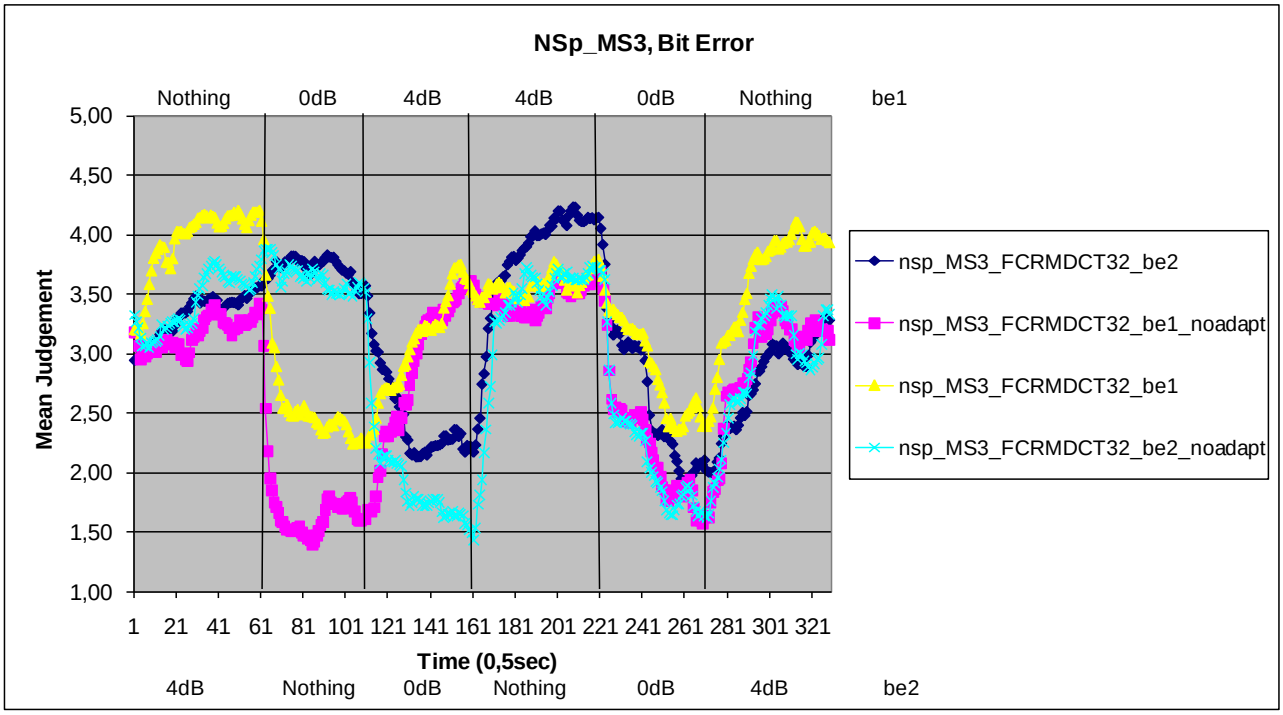


Figure 8: Mean instantaneous responses obtained for 2 packet loss conditions, obtained with the speech sequence NSp_MS3.

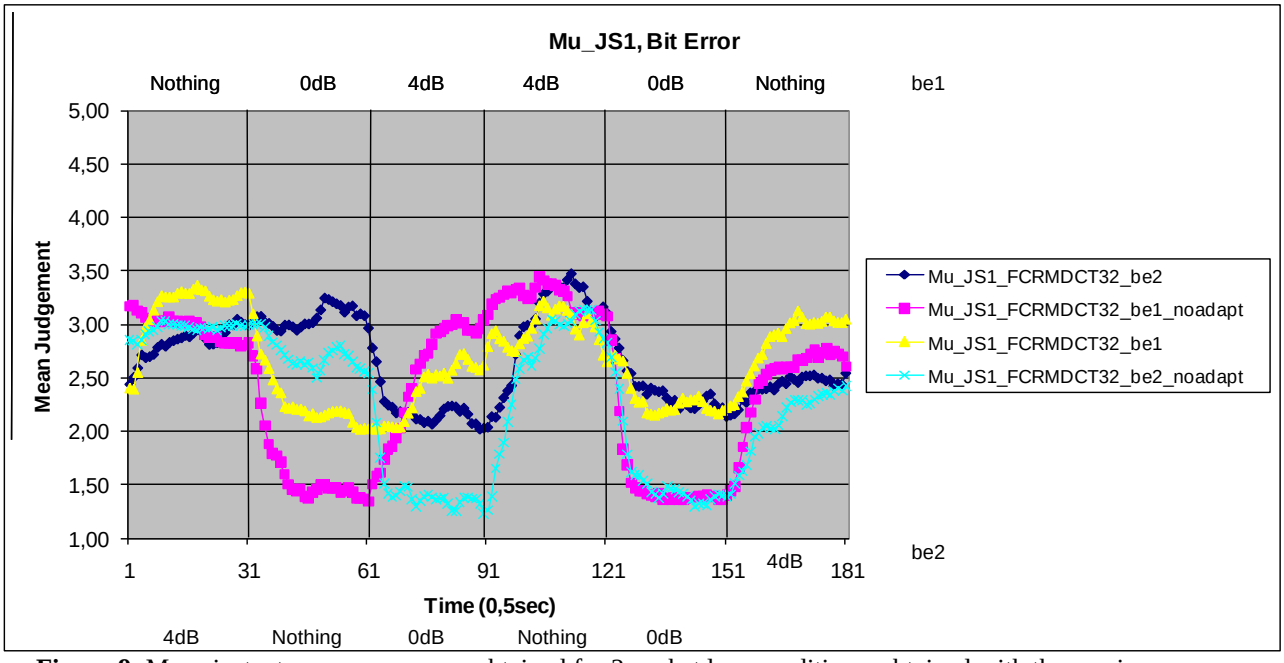


Figure 9: Mean instantaneous responses obtained for 2 packet loss conditions, obtained with the music sequence Mu_JS1.

IV. “QUALITY” TESTS

1. Test Sequences

The test items were chosen to be realistic types of audio excerpts as much as possible, keeping in mind that they must remain as critical as possible as well (that means that transparency is not often achieved by famous encoders when encoding those audio sequences). A set of items was chosen for the test, and another one, smaller, was chosen for the training phase of the test. The main items are listed in table 6. The training items are listed in table 7.

Category	Description	Origin	Name
Speech	English, Female	EBU-SQAM Track 49 (0s - 7.6s)	SpEn_FS1_Ref.wav
Speech	English, Female	NTT AT CD1 Track 20 Right (1mn30s-1mn33s)	SpEn_FS2_Ref.wav
Speech	English, Male	NTT AT CD1 Track 20 Left (0-3s)	SpEn_MS1_Ref.wav
Speech	English, Male	EBU-SQAM Track50 (0s-7.5s)	SpEn_MS2_Ref.wav
Music	English Female singer – Jazz music	Orange Database	Mu_PS1_Ref.wav
Music	English Male singer – Soft pop music	3GPP m_po_x_4_org.wav	Mu_PS2_Ref.wav
Music	Soft pop music (Eric Clapton – Leila)	3GPP	Mu_PS3_Ref.wav
Speech + Music	Male speech + music	3GPP sbm_sm_x_1_org.wav	SpMu_MS1_Ref.wav
Speech + Music	Male speech + noise and music (trailer)	3GPP som_fi_x_2_org.wav	SpMu_MS2_Ref.wav
Speech + Music	Female speech + Music	3GPP som_fi_x_4_org.wav	SpMu_FS1_Ref.wav
Noisy speech	English male speeches + noises	3GPP s_no_2t_1_org.wav	NSp_MS1_Ref.wav
Noisy speech	English male speech + noise	3GPP s_no_ft_3_org.wav	NSp_MS2_Ref.wav

Table 6: Main test items

Category	Description	Origin	Name
Speech	English, Male	NTT AT CD1 Track 21 Left (1mn38s-1mn42ss)	SpEn_MS3_Ref.wav
Music	Orchestra	EBU-SQAM Track66 (0s-6s)	Mu_COS2_Ref.wav
Speech + Music	English female speaker + guitar	3GPP sbm_sm_x_5_org.wav	SpMu_FS3_Ref.wav
Noisy speech	English Male talkers + noise	3GPP s_no_2t_1_org.wav	NSp_MS4_Ref.wav

Table 7: Training test items

The duration of the items was ranging from 4 seconds to 15 s. Fade in and fade out were generated when necessary. The loudness alignment of all sequences was done following the processing test plan [2].

2. Test Methodology

The test methodology MUSHRA was used. MUSHRA stands for MUlti Stimuli with Hidden Reference and Anchor points. This is a method dedicated to the assessment of intermediate quality. It has been recommended at the ITU-R under the name BS.1534 [8].An important feature of this method is the inclusion of the hidden reference and bandwidth limited anchor signals. The chosen anchor point was the band-limited signal with cut-off frequencies of 3.5 kHz (mandatory). The generation of the anchors sequences meets the requirements of the processing test plan [2].

3. User Interface

The MUSHRA method has the advantage of displaying all stimuli for one test item at a given bit-rate or configuration at the same time. The subjects were therefore able to carry out any comparison between them directly.
A screenshot of an example of the MUSHRA user interface from CRC-Ottawa-Canada (SEAQ-used at Orange labs) is shown in figure 10 A screenshot of an example of the MUSHRA user interface used at Ericsson is shown in figure 11. The buttons represent all the configurations/codecs under test including the hidden reference and the anchor signal, and the reference, which is also displayed on the left as "REF". Above each button, with the exception of the "REF" one, a slider is used to grade the quality of the test item according to the reference signal quality all along a continuous quality scale:

[0-19]: Bad
[20-39]: Poor
[40-59]: Fair
[60-79]: Good
[80-100]: Excellent

For each of the test items, the signals under test were randomly assigned, with a different assignment for each subject. In addition, the test items were randomised for each subject within a session to avoid sequential effects.

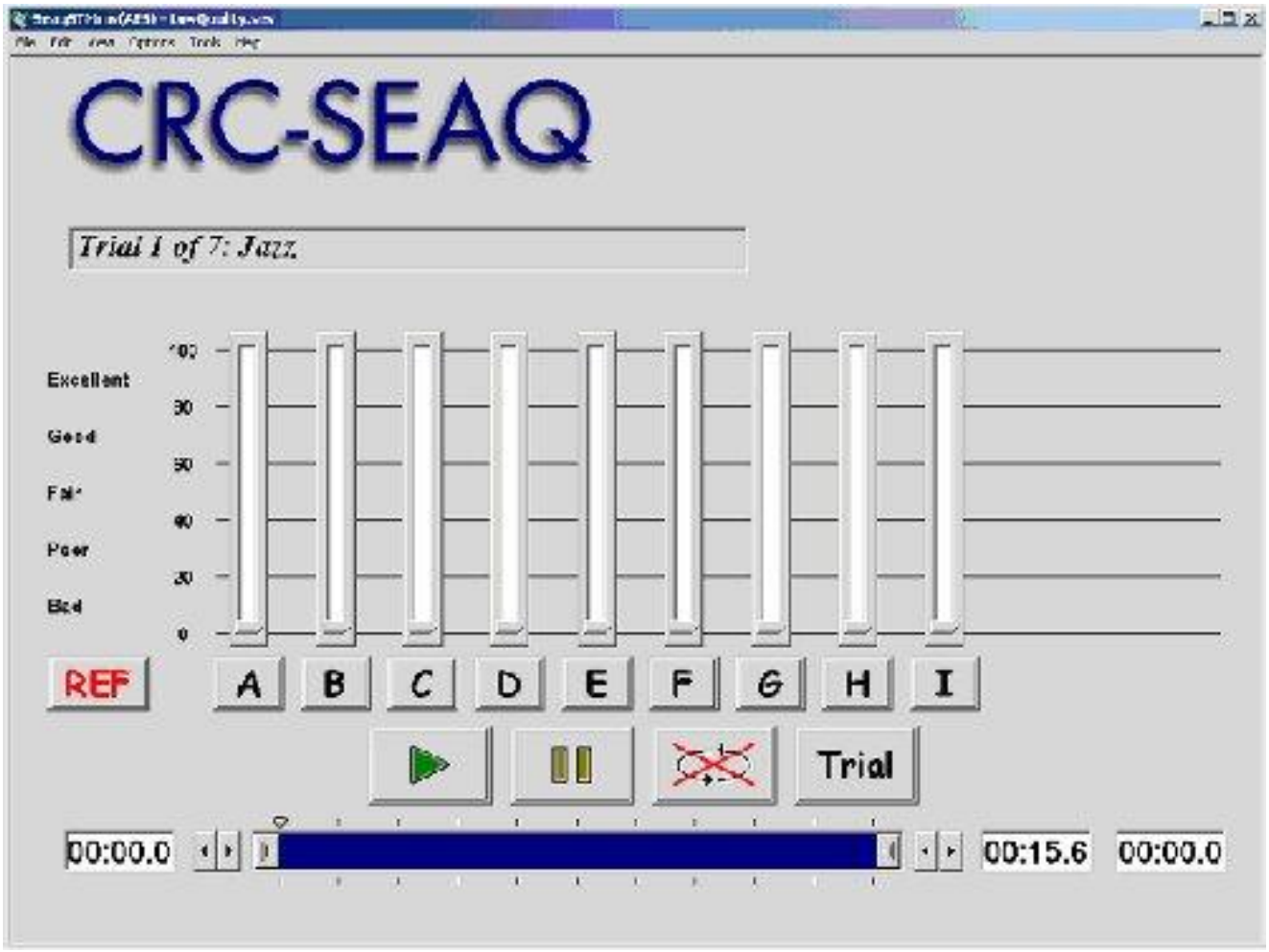


Figure 10: CRC-SEAQ MUSHRA Software

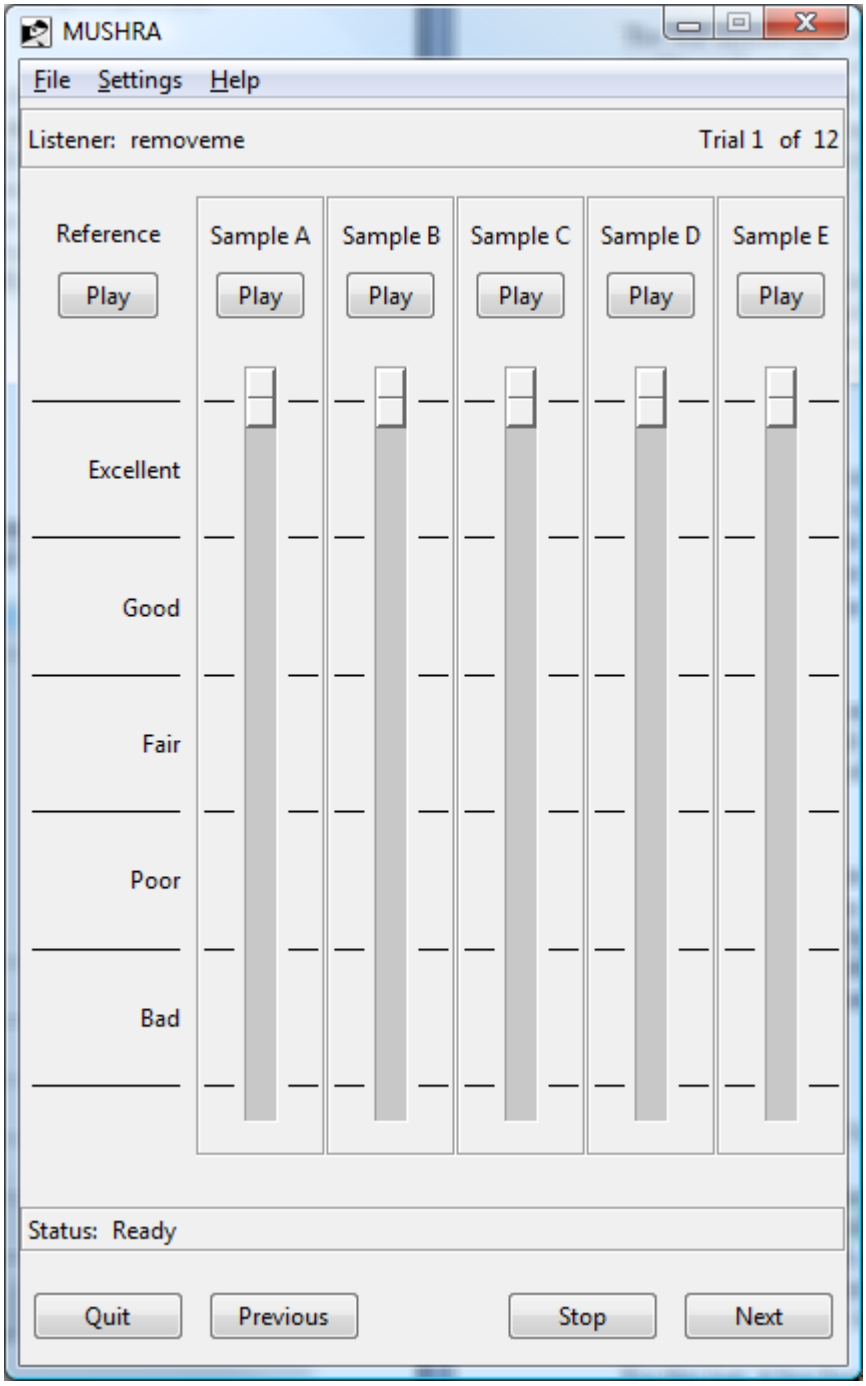


Figure 11: MUSHRA Interface used at Ericsson.

4. Training Phase

Each listener had a period of training of about 15 min, in order to get familiar with the test methodology and software and with the kind of quality they have to assess. This was as well an opportunity to adjust the playback level that would then remain constant during the test phase.

5. Test instructions and duration

The test instructions explain to the listeners how the software works, what they will listen to (briefly), how to use the quality scale and how to score the different excerpts. An example is given in the quality assessment plan [2]. This is as well an opportunity to mention the fact that there is a hidden reference signal to score and consequently, there should be at least one score equal to 100 per excerpt. This will be used later on in the rejection process of listeners.

The test duration was different according the test. At Orange labs the duration was timed as follows

- 1h15 on the average for the high bit-rate test (32kbps)

- 1h on the average for the medium bit-rate test (24kbps)
- 45mn on the average for the low bit-rate test (14kbps)

Such a gap is because of the quite high quality encountered at 32kbps. Listeners took more time when scoring, specially trying to find out the hidden reference.
Every 20 min, the listener was asked to rest a bit by breathing some fresh air. The test instructions and training phase were not included in this time schedule.

6. Listening panel

The listening panel at Orange labs consisted altogether of 7 to 9 subjects (according the test), most of them experienced in audio but not professionally involved. There were:

- 11 listeners for the 14kbps test ;
- 11 listeners for the 24kbps test ;
- 12 listeners for the 32kbps test ;

At Ericsson the number of listeners was 6 for the 14.25 kbps and 32 kbps tests and 7 for the 24 kbps test. All listeners are professionally involved in audio signal processing research.
At Nokia the number of listeners was 6 for the 14.25 kbps and 24 kbps tests and 5 for the 32 kbps test. All listeners were hired outside Nokia and were naïve regarding speech and audio processing. Listeners were, however, familiar with the subjective testing and MUSHRA methodology particularly.

7. Listening Conditions

At Orange the tests were performed on the headphone STAX Signature SR-404 (open model) and its amplifier SRM-006t. The test items were stored on a Windows 2k workstation. The digital sound was played through the PC board Digigram VX 222 and converted with a 24 bits DAC (3Dlab DAC 2000). At Ericsson Sennheiser HD250 headphones were used connected to an Edirola UD-25 A/D converter. At Nokia, Sennheiser HD580 headphones were used.

8. Statistical analysis

The statistical analysis method described in the MUSHRA specifications was used to process the test data. The results are presented as mean grades and 95% confidence intervals of the standard deviation of the mean.
Experience has shown that the scores obtained for different test sequences are dependent on the criticality of the test material used. In order to provide a more complete understanding of codecs performances results are presented separately for different test sequences rather than only as aggregated averages across all the test sequences used in the assessment.

9. Post-screening of subjects

The post screening of subjects is performed in different steps.

Firstly, due to the fact that "low" quality is tested, a subject should be able to easily identify the hidden reference signal from the coded versions. That means that listeners that are not able to isolate and score "100" for the hidden reference signal (with a pre-defined error) should be discarded. Usually, people who have scored the hidden reference above "90" are not rejected as long as they haven't scored the band limited reference signals higher than the reference one. This value (90) can be adjusted according to the overall quality of the test: the higher the quality, the lower the threshold.
For this test, it has been decided not to set any threshold as there were too many listeners involved.

Then, a subject should be able to give a grade close to the grade given by the majority of the subjects. Some subjects may have difficulties in reliably assessing the audio quality of the provided excerpts. That is reflected in scores that differ substantially from the average scores given in the test by all subjects.
The easiest way to measure the inconsistencies of an individual subject compared to the mean result is to calculate both the correlation coefficient and the mean square error.

The correlation coefficient $\rho_{x,y}$ is used to determine the relationship between 2 sets of data. It is calculated as follows:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x.\sigma_y} \qquad \text{With} \quad -1 \le \rho_{x,y} \le 1$$

$$\text{Where} \quad Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$$\sigma_x = \sqrt{\frac{1}{n}.\sum_{i=1}^{n}(x_i - \mu_x)^2}$$ (Respectively with the Y set)

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$ (Respectively with the Y set)

n is the number of conditions * number of audio excerpts
X is the set of n scores given by the listener x
Y is the set of the n average scores over all the listeners

It has been decided to discard people whose correlation coefficient $\rho_{x,y}$ is below 0.8.

At Orange the whole process leads to discard:

- Medium bit-rate test : 1 listener (over 11)
- High bit-rate test : 2 listeners (over 12)

At Ericsson no listeners had to be discarded as all listeners reached a correlation coefficient larger than 0.9.
At Nokia no listeners were discarded.

10.Results

1. Low bit rate: 14.25kbps

a. Orange Labs

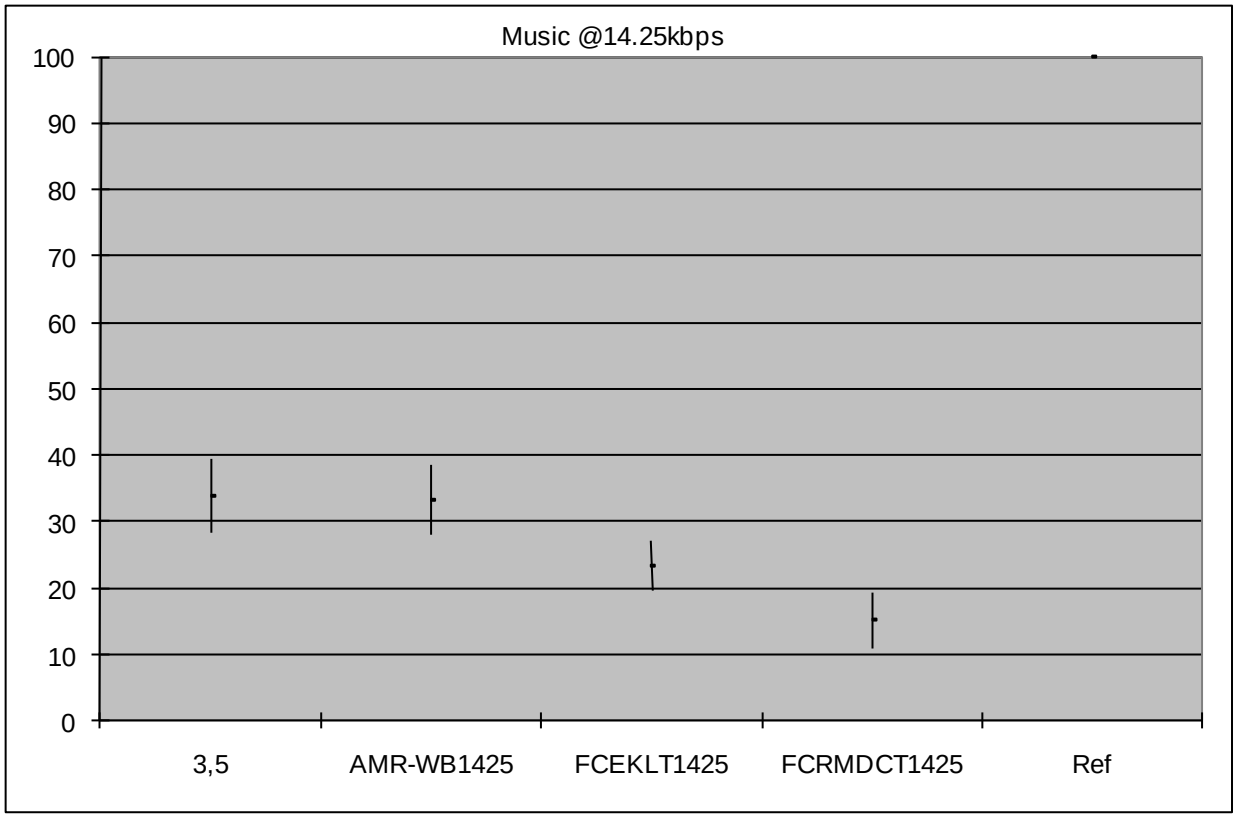


Figure 11: Results on the 3 Music items (14.25kbps – Orange Labs)

Figure 11 shows the results on the 3 music items over all listeners at 14kbps. What is noticeable is that the overall quality is very low, whatever the tested codec. The 2 *FlexCode* codecs have been judged "bad" (0-20), even if

the FCEKLT performs a bit better than the FCRMDCT, in between “bad” and “poor” (20-40). But both of them are scored below the 3.5kHz anchor.

The anchor codec AMR-WB has been scored around 34, poor quality (20-40), a bit above the anchor 3.5kHz.

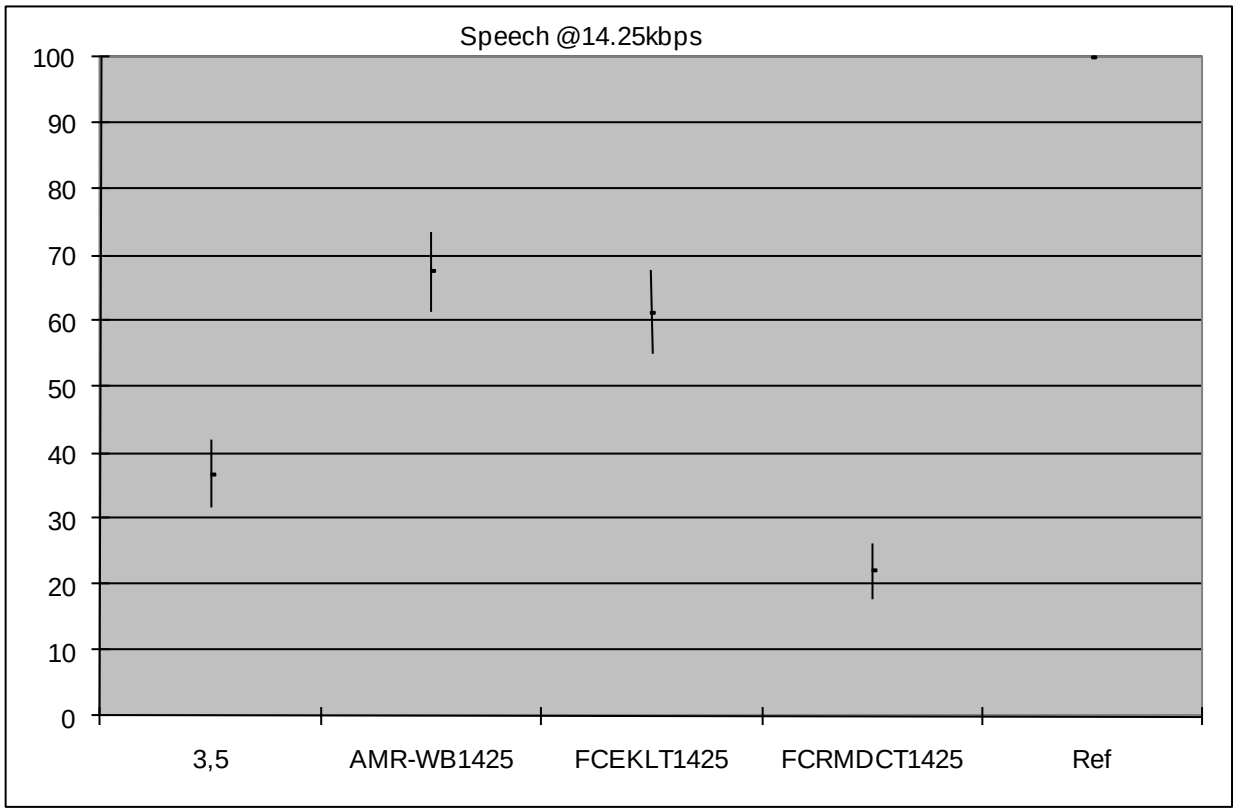


Figure 12: Results on Speech items (4 excerpts) (14.25kbps – Orange Labs)

Figure 12 shows the results on the 4 speech items over all listeners. We can note the intermediate “fair” / “good” (around 60) behaviour in quality of the FCEKLT: its average score is quite close to the one of the AMR-WB, judged “good” (60-80).

The FCRMDCT is scored “poor” (20-40), below the anchor 3.5kHz scored “poor” (20-40).

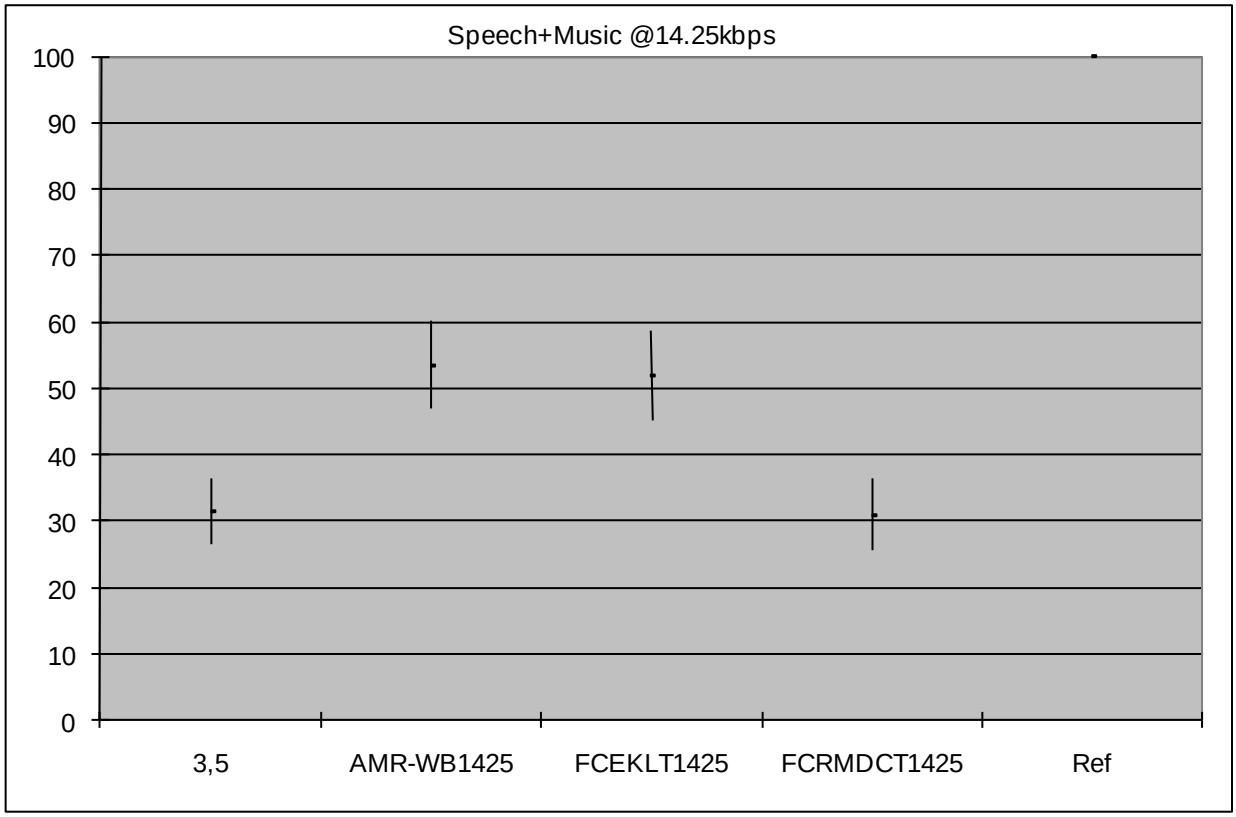


Figure 13: Results on Speech + Music items (3 items) (14.25kbps – Orange Labs)

The AMR-WB and FCEKLT codecs are scored "fair" (40-60). The FCRMDCT is judged "poor" (20-40), the same as the 3.5kHz anchor.

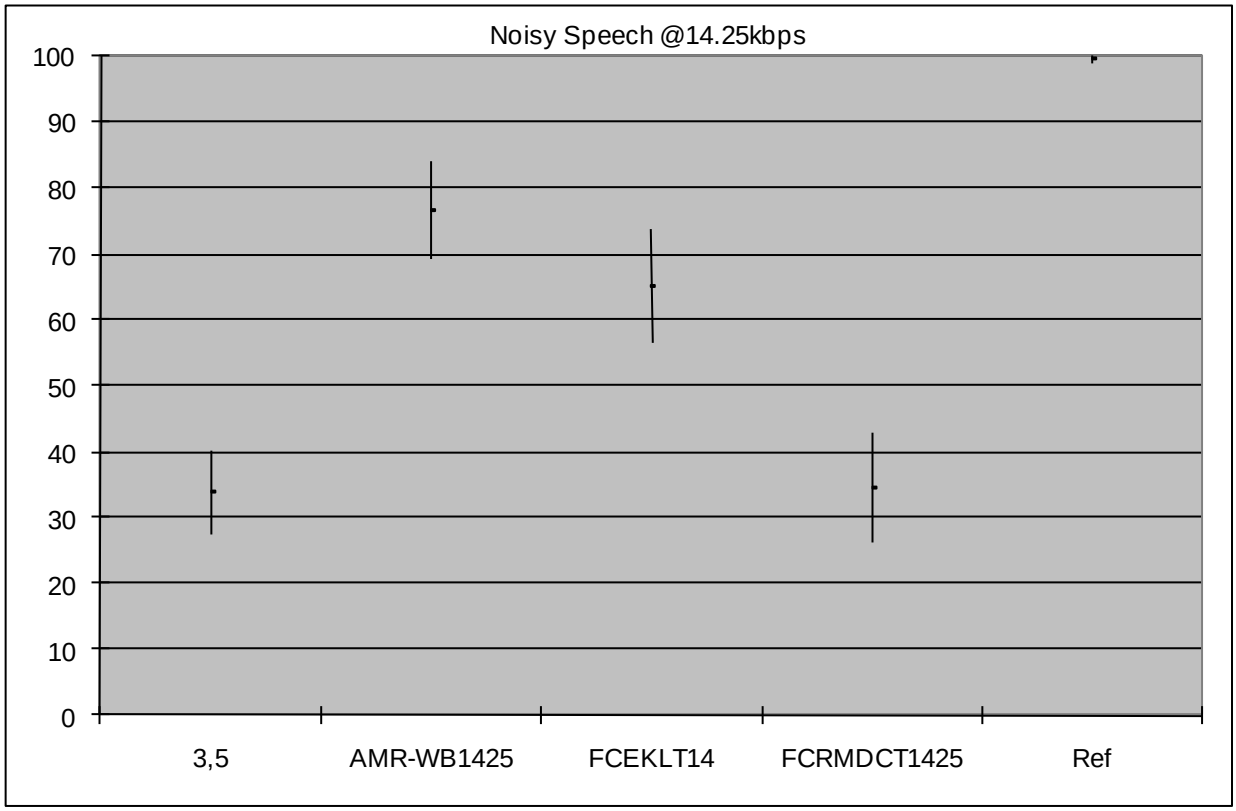


Figure 14: Results on Noisy speech items (2 items) (14.25kbps – Orange Labs)

The FCEKLT and AMR-WB codecs have been judged “good” (60-80), the anchor codec being better than the tested one. The FCRMDCT is judged “poor” (20-40), the same as the 3.5kHz anchor.

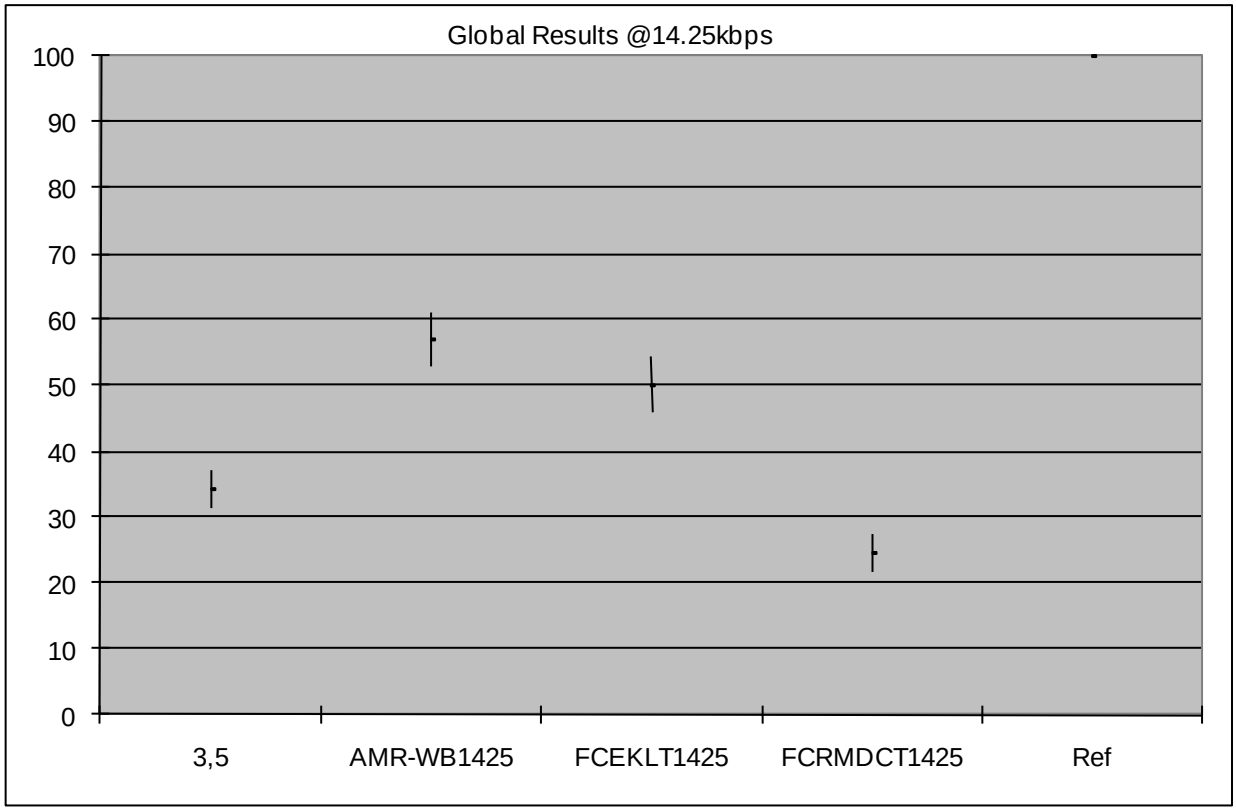


Figure 15: Results on all 12 items (14.25kbps – Orange Labs)

Figure 15 shows the results over the 12 audio items for all listeners.

The FCRMDCT is judged “poor” (20-40) below the 3.5 kHz anchor which quality is judged “poor” as well. The FCEKLT is scored “fair” (40-60) below the AMR-WB scored fair as well.

Student	3,5	AMR-WB1425	FCEKLT1425	FCRMDCT1425	Ref
3,5	1,00				
AMR-WB1425	0,00	1,00			
FCEKLT1425	0,00	0,02	1,00		
FCRMDCT1425	0,00	0,00	0,00	1,00	
Ref	0,00	0,00	0,00	0,00	1,00

Table 8: Student Test on all 12 items (14.25kbps – Orange Labs)

Table 8 gives the results of a Student T test.

Figures calculated by a Student T test are the probability that two compared configurations are significantly different or not (intersection between a line and a column).

In our case, this test is used to observe whether the quality of a coder or specific configuration is significantly different from that of the other coders / configurations. The following assumptions were made in order to calculate table 8:

- The Student T test uses the bilateral distribution ;
- The T test was done over two set of samples with different standard deviation.

A number higher than 0.05 means that the two compared configurations are not statistically different.

Consequently, figures in table 8 indicate that evaluated configurations are statistically different.

b. Ericsson

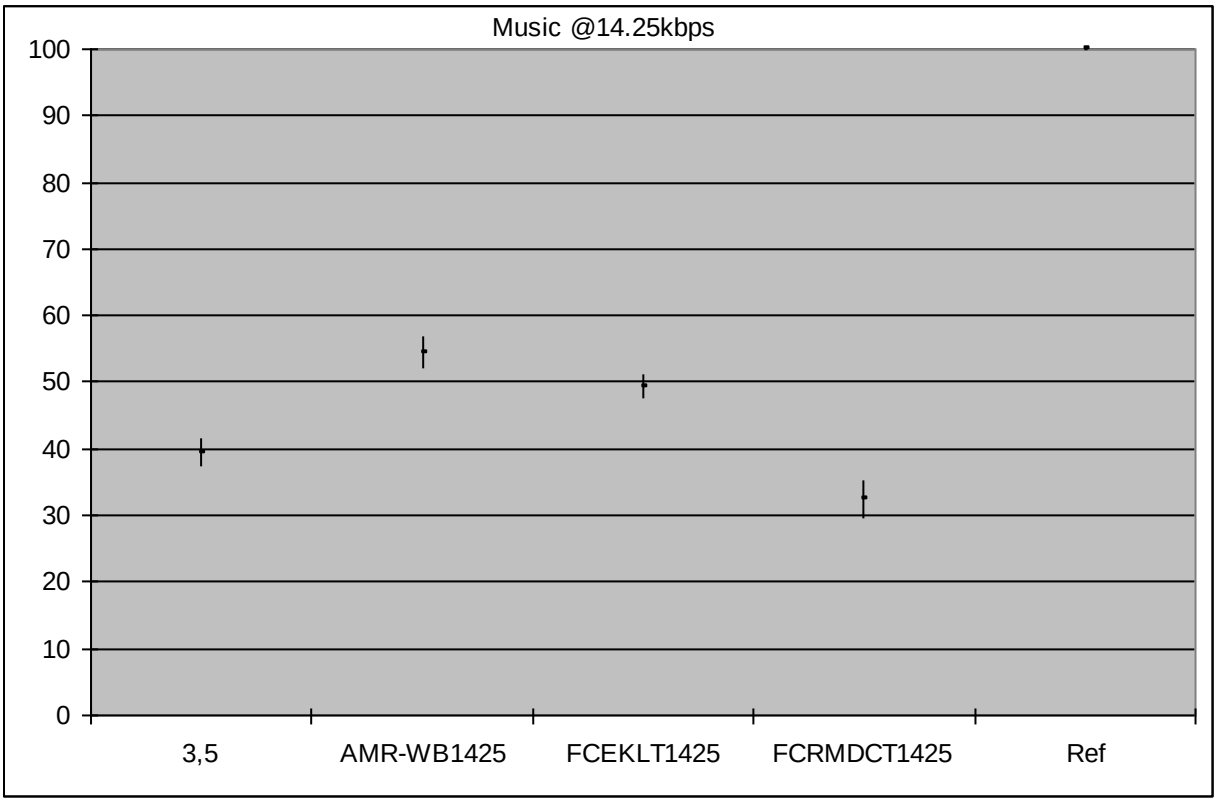


Figure 16: Results on the 3 Music items (14.25kbps – Ericsson)

Figure 16 shows the results for the 3 music items over all 6 listeners for 14kbps. The same trend as for the Orange results is observed. However, all scores are higher than the scores recorded at Orange Labs. This is true for all the 14.25 kbps shown in this section. A further difference is that the FCEKLT performance is judged closer to the AMR-WB performance.

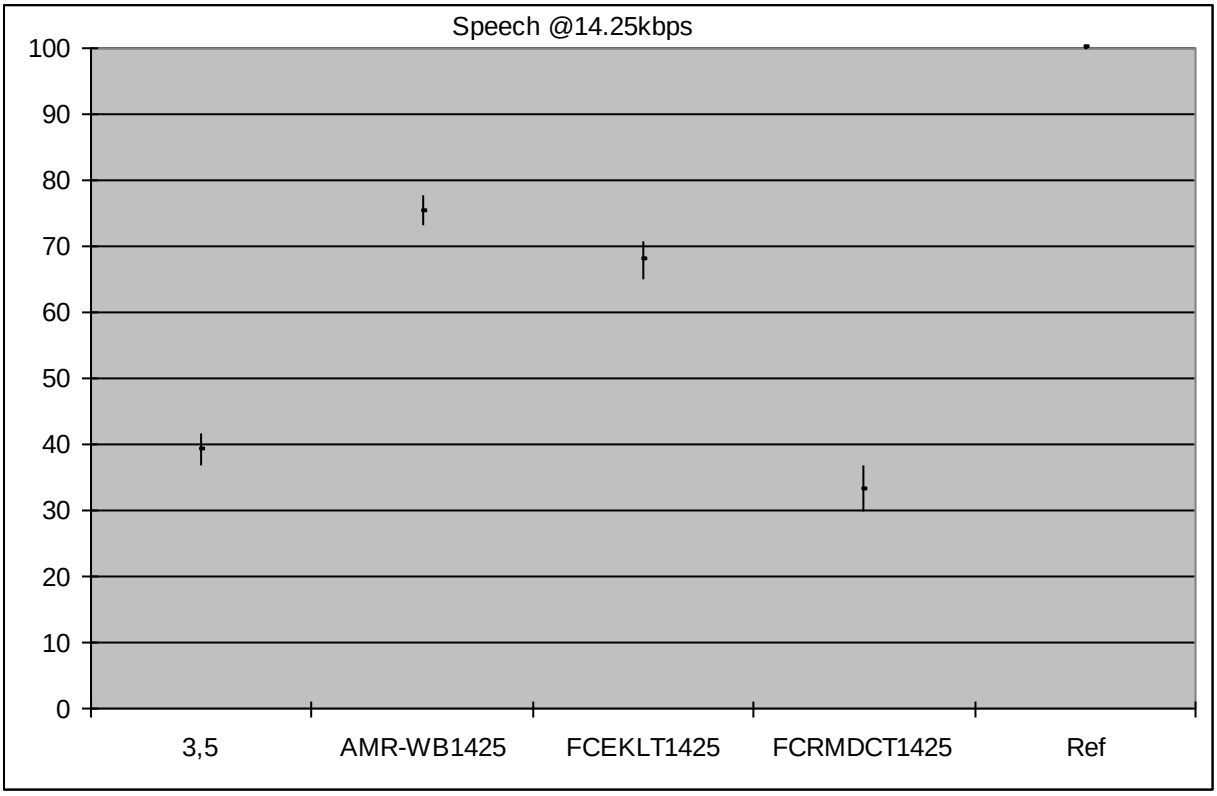


Figure 17: Results on Speech items (4 excerpts) (14.25kbps – Ericsson)

Figure 17 shows the results for the 4 speech items. Here the trend is similar to the one reported at Orange Labs.

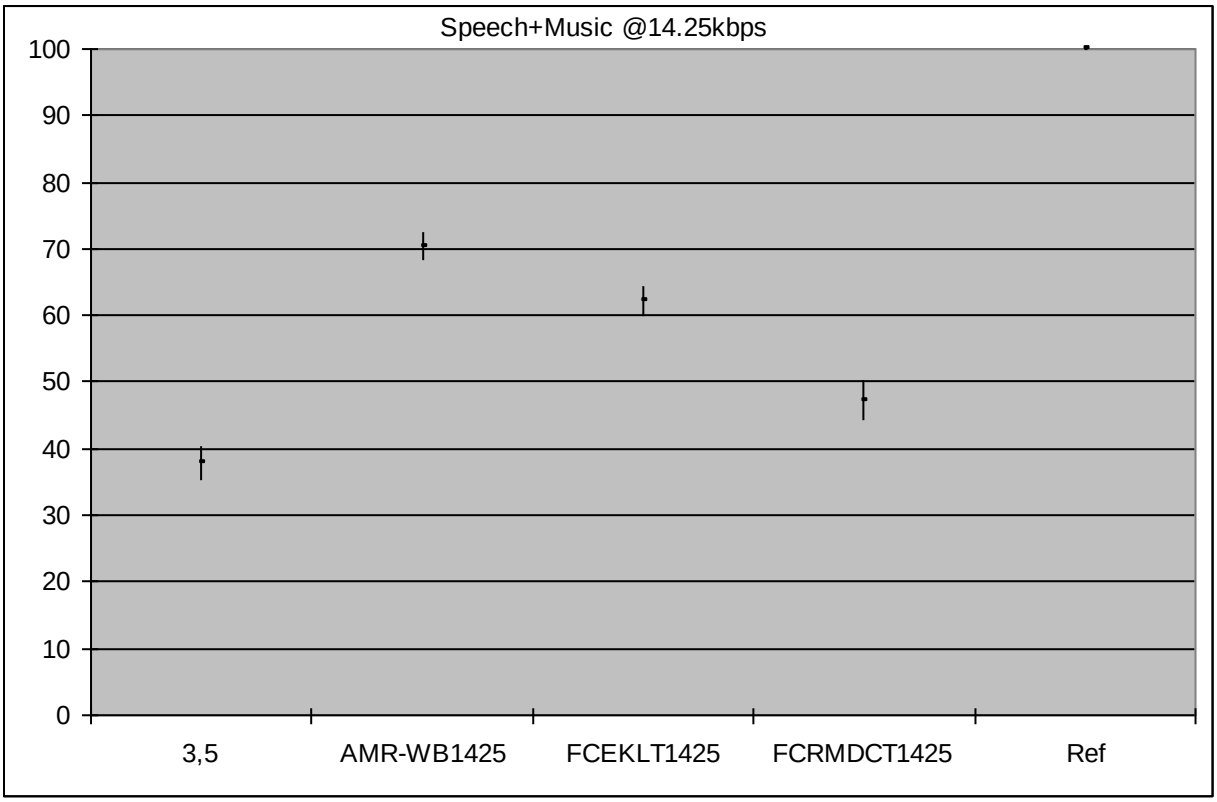


Figure 18: Results on Speech + Music items (3 items) (14.25kbps – Ericsson)

Figure 18 shows the results for the speech+music items. The ordering of the codecs is similar as before. However, the FCRMDCT codec is scored clearly better than the 3.5kHz anchor.

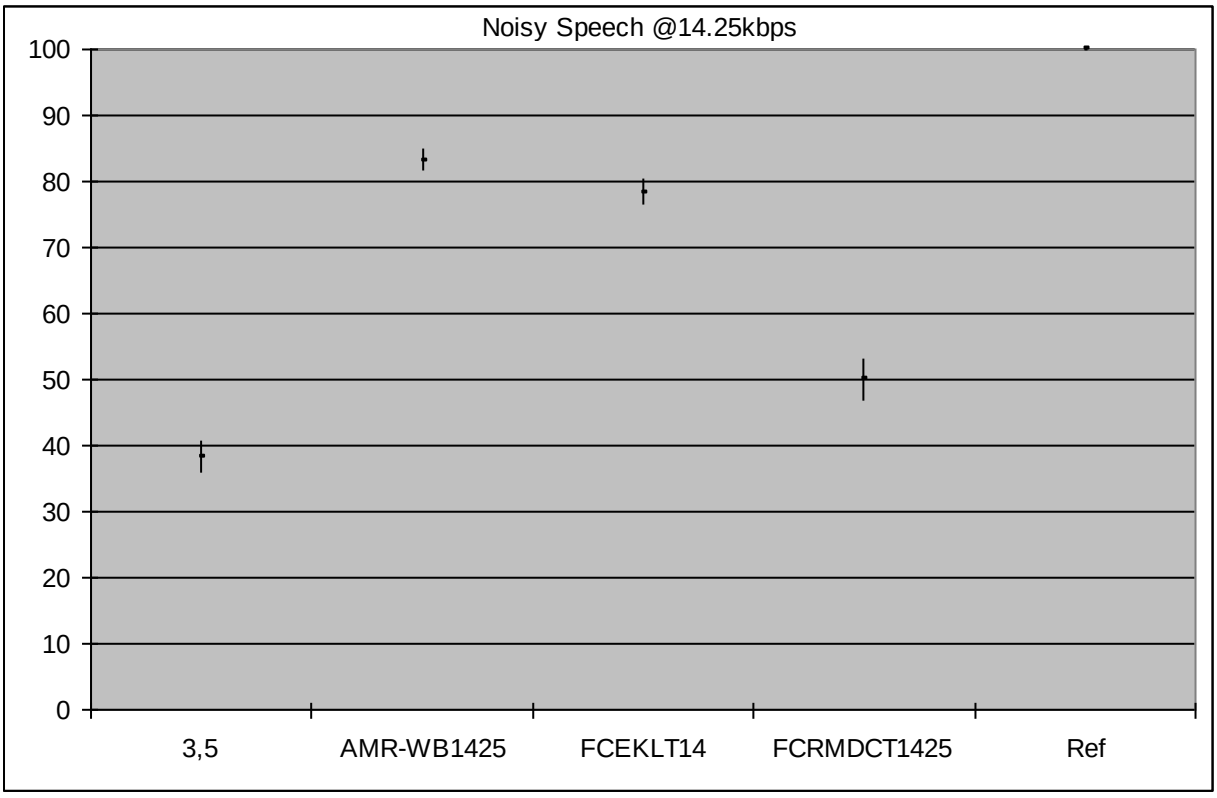


Figure 19: Results on Noisy speech items (2 items) (14.25kbps – Ericsson)

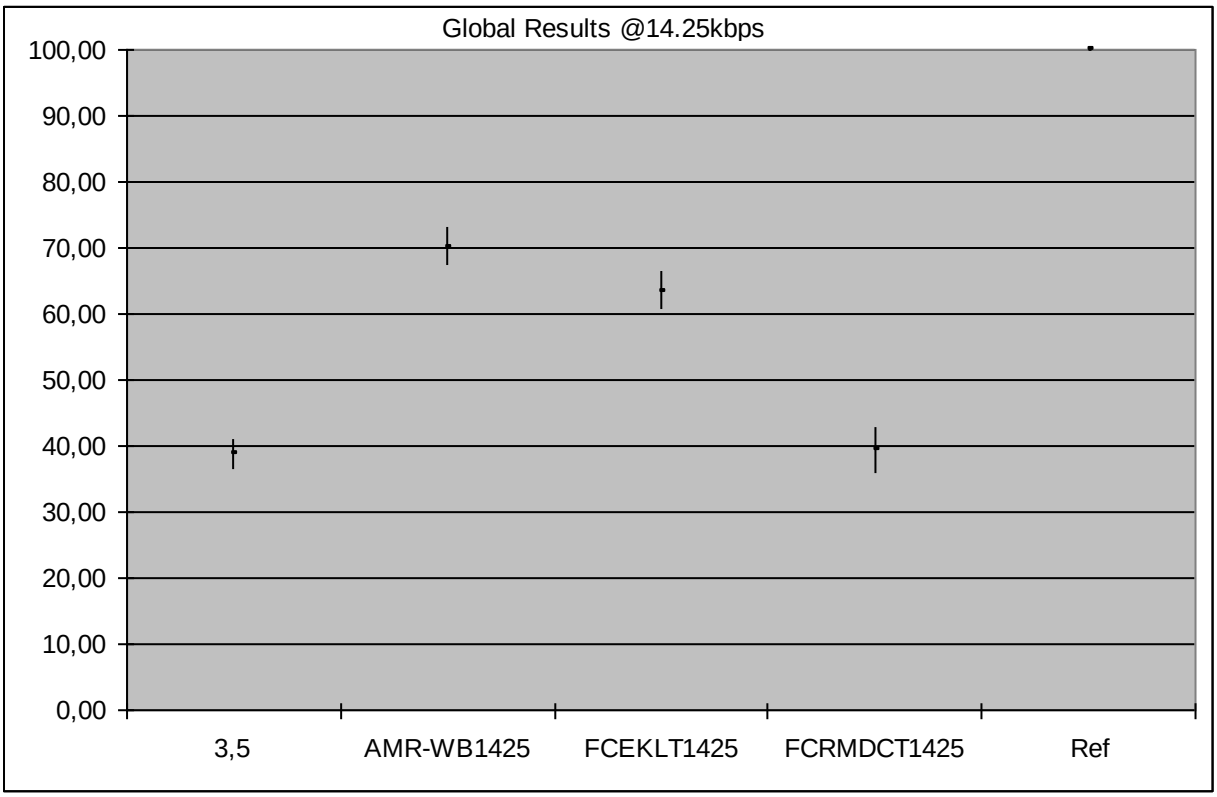


Figure 20: Results on all 12 items (14.25kbps – Ericsson)

Figure 20 shows the scores for all 12 items from all 6 listeners. The FCEKLT codec scores about 7 MUSHRA points worse than the AMR-WB codec. The FCRMDCT codec is close to the 3.5 kHz anchor and about 23 MUSHRA points below the FCEKLT codec.

Student	3,5	AMR-WB1425	FCEKLT1425	FCRMDCT1425	Ref
3,5	1,00				
AMR-WB1425	0,00	1,00			
FCEKLT1425	0,00	0,01	1,00		
FCRMDCT1425	0,82	0,00	0,00	1,00	
Ref	0,00	0,00	0,00	0,00	1,00

Table 9: Student Test on all 12 items (14.25kbps – Ericsson)

The figures from the student test shown in Table 9 indicate a difference between the AMR-WB, FCEKLT, and FCRMDCT codec. The FCRMDCT and 3.5kHz can not be separated with 95% confidence.

c. Nokia

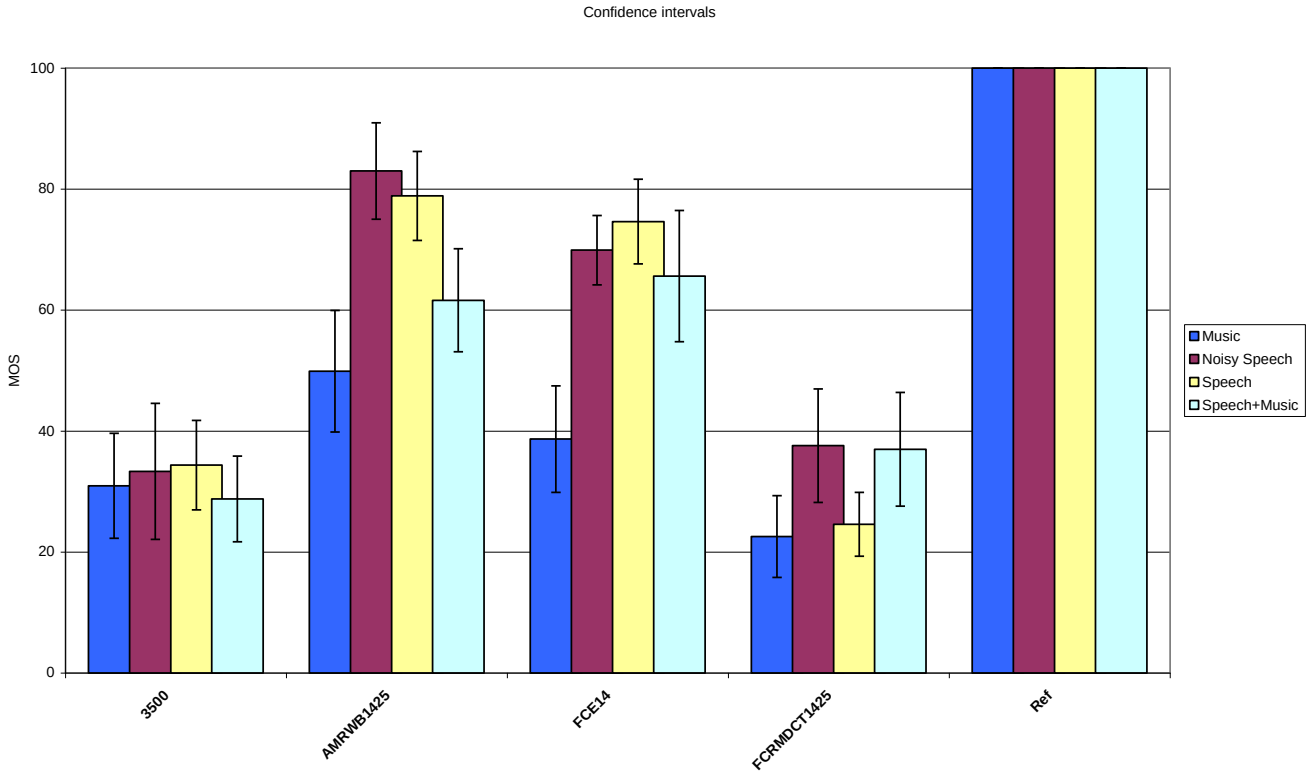


Figure 21: Results on the 3 Music items (14.25kbps – Nokia)

Figure 21 shows the results for the 3 music items 4 speech items, 3 speech+music and 2 noise speech items over all 6 listeners for 14kbps. The results show same trend as for the Orange and Ericsson results while the absolute scores a bit higher than in the Orange site.

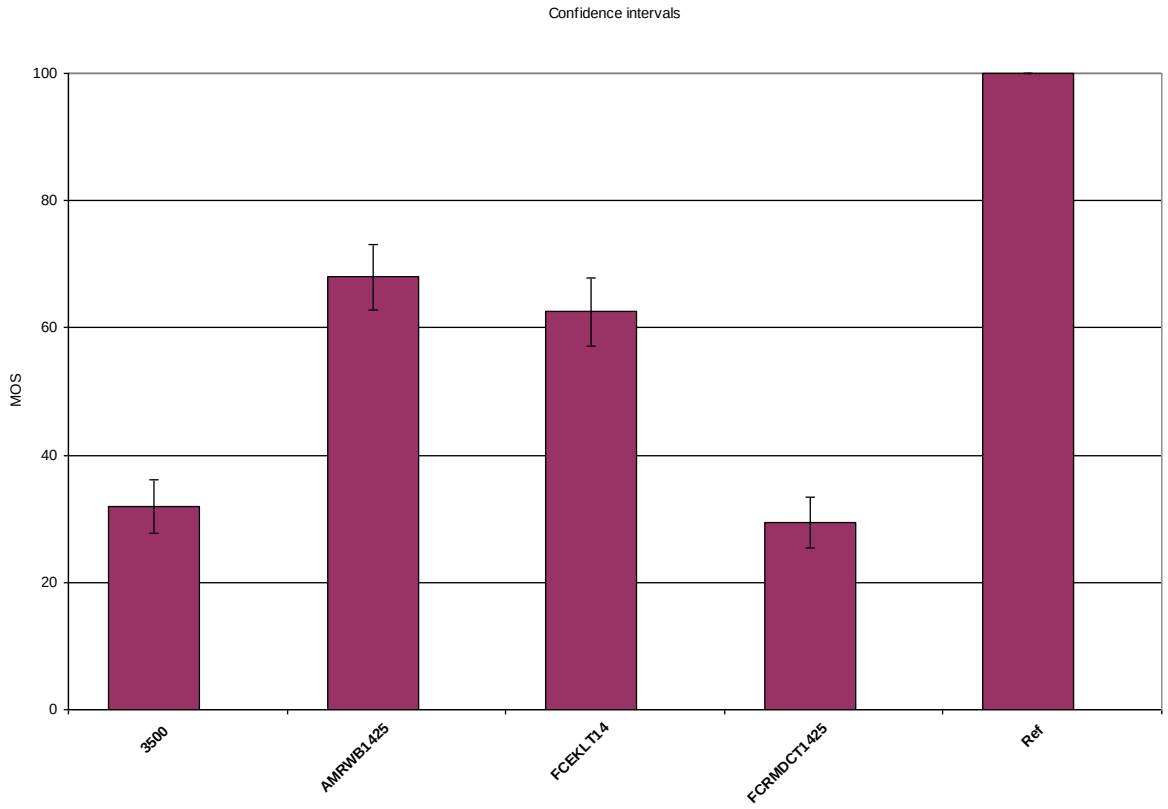


Figure 22: Results on all 12 items (14.25kbps – Nokia)

Figure 22 shows the scores for all 12 items from all 6 listeners. The FCEKLT codec scores about 6 MUSHRA points worse than the AMR-WB codec.

	3500	AMRWB1425	FCE14	FCRMDCT1425	Ref
3500					
AMRWB1425	BT				
FCE14	BT	EQ			
FCRMDCT1425	EQ	WT	WT		
Ref	BT	BT	BT	BT	

Table 10: Student Test on all 12 items (14.25kbps – Nokia) (BT – better than, Eq – equal, WT – worse than)

The results from the student test shown in Table 10 indicate a difference between the AMR-WB, FCEKLT, and FCRMDCT codec. The FCRMDCT and 3.5kHz can not be separated with 95% confidence.

2. Medium bit rate: 24kbps

a. Orange Labs

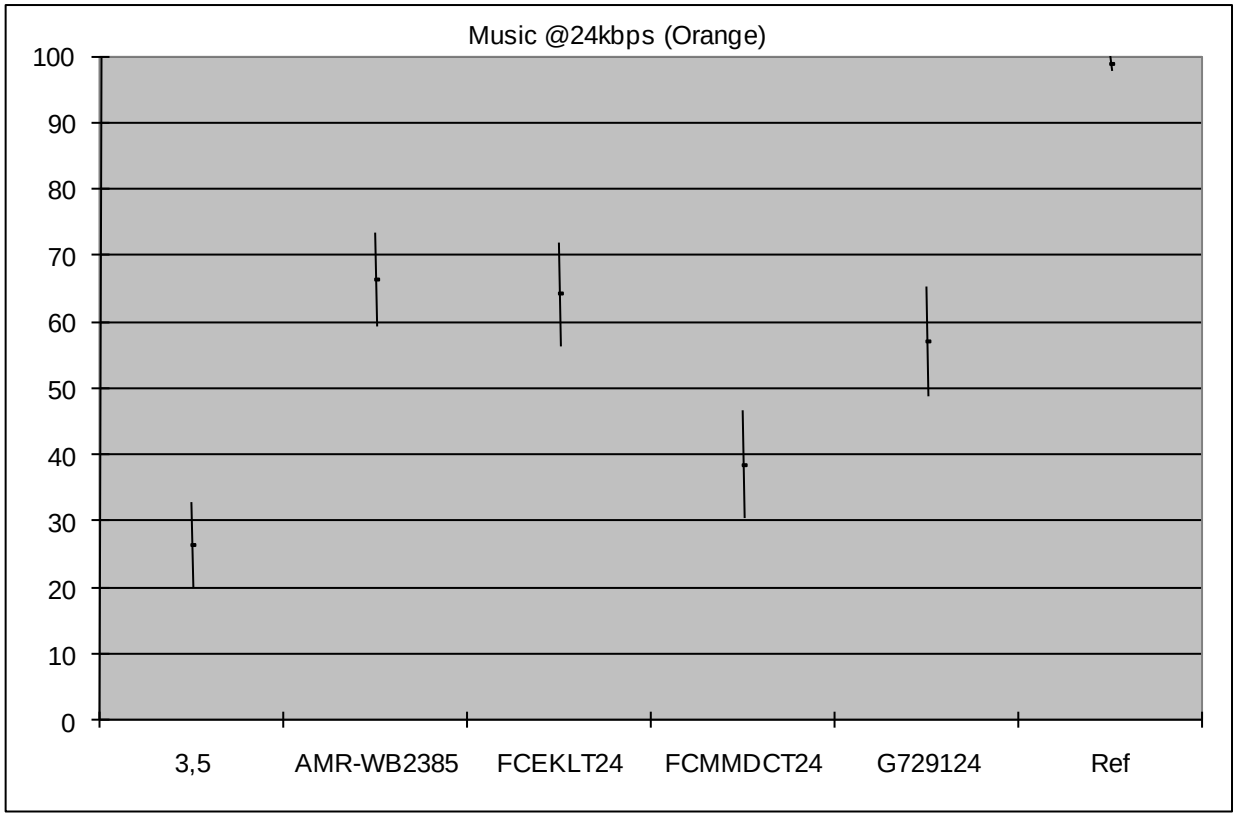


Figure 23: Results on Music items (3 items) (24kbps – Orange Labs)

Figure 23 shows the results on the 3 music items over all listeners. The first remark is that on such items, the FCEKLT has been scored the same as the AMR-WB "good" quality (60-80). A bit below, the G729.1 is judger "fair" (40-60). The FCRMDCT has been scored "poor" (20-40), near the "fair" quality and above the 3.5 kHz anchor in the same quality range.

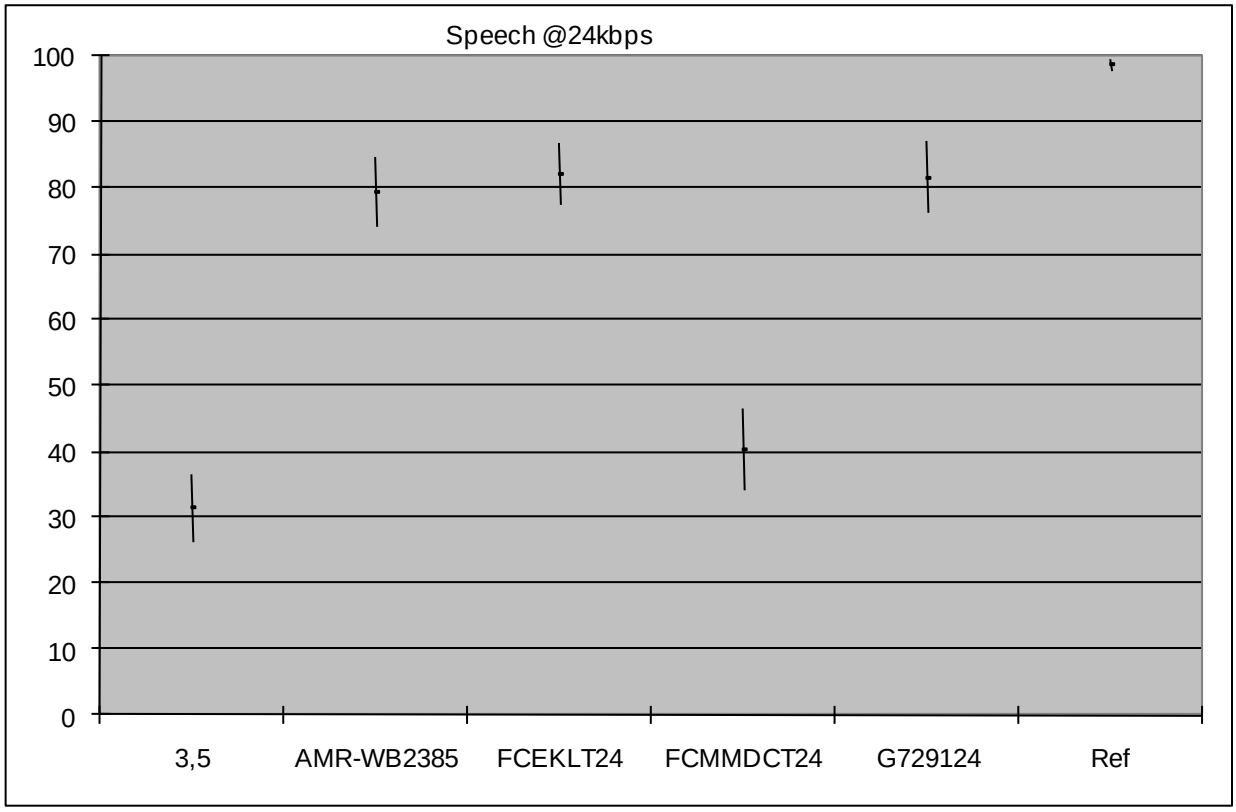


Figure 24: Results on Speech items (4 items) (24kbps – Orange Labs)

Figure 24 represents the results obtained on all 4 speech items over all the listeners at 24kbps bit rate. The FCEKLT has been scored excellent (80-100), the same as the G729.1, and a bit above the AMR-WB judged "good" (60-80) but quite close to the "excellent" quality border. The FCRMDCT is judged "fair" (40-60), above the 3.5 kHz anchor scored "poor" (20-40).

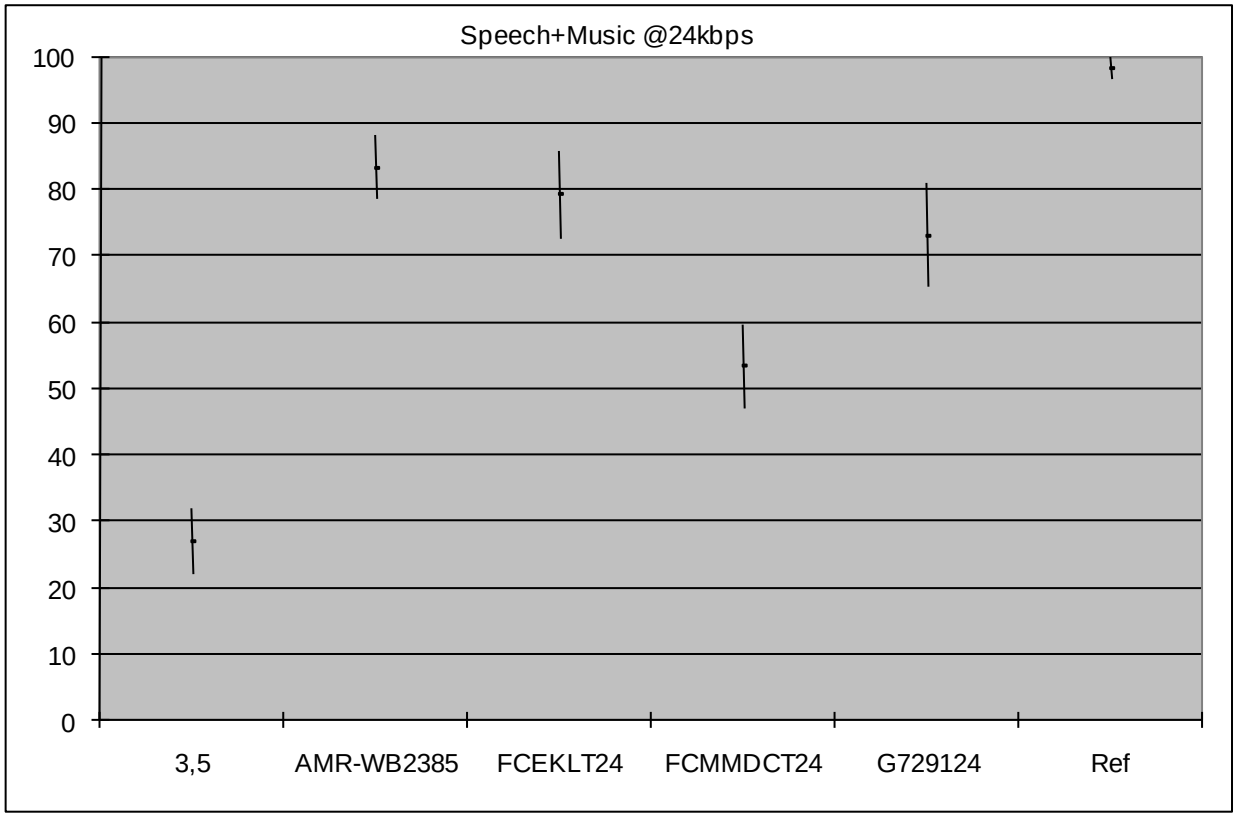


Figure 25: Results on Speech + Music items (3 items) (24kbps – Orange Labs)

Figure 25 shows the results on the 3 speech and music items over all listeners. The FCEKLT has been scored 78 that is "good" quality (60-80), just below the AMR-WB judged "excellent" (80-100) with an average score of 82. A bit below, the G729.1 is judger "good". The FCRMDCT has been scored "fair" (40-60) above the 3.5 kHz anchor (scored "poor" (20-40)).

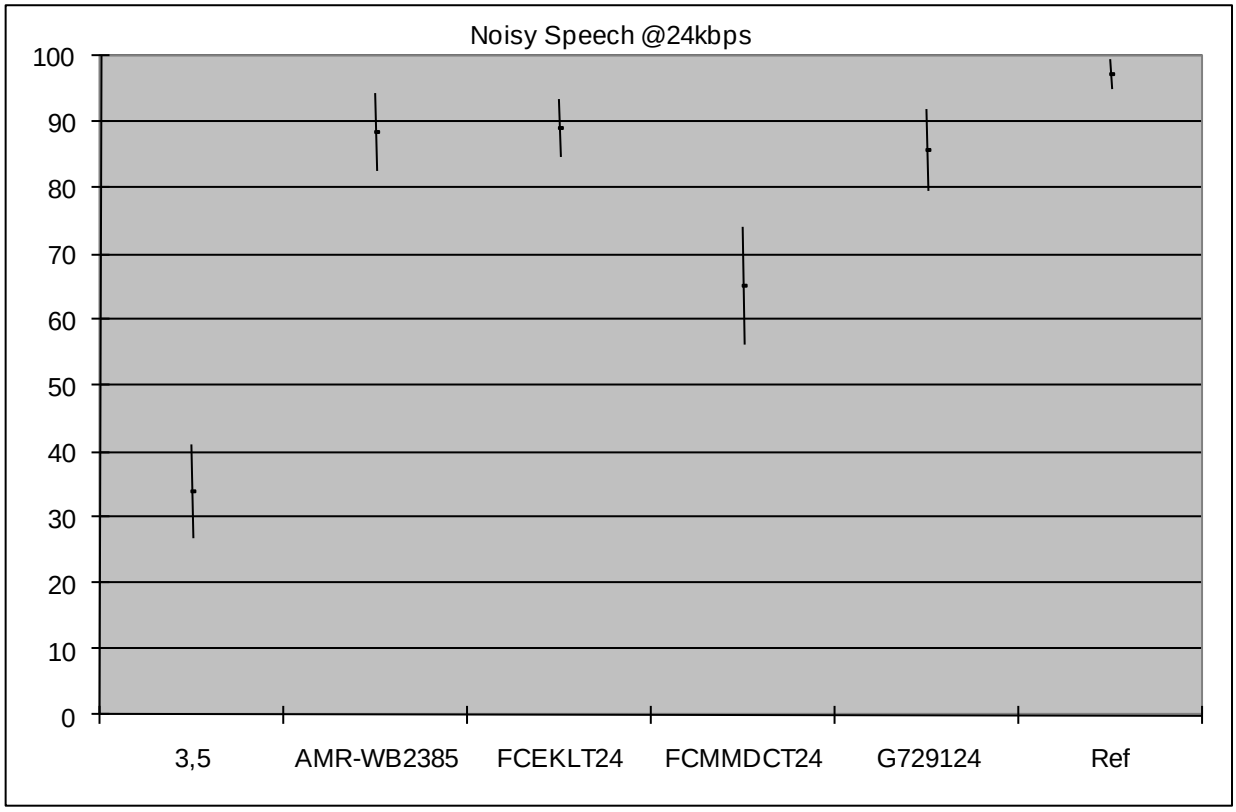


Figure 26: Results on Noisy speech items (2 items) (24kbps – Orange Labs)

Figure 26 represents the results for the 2 noisy speech items over all the listeners at 24kbps. The AMR-WB is scored "excellent" (80-100), the same as the FCEKLT codec and a bit above the G729.1 codec scored excellent as well. The FCRMDCT is judged "good" (60-80), above the 3.5 kHz anchor score ("poor" (20-40)).

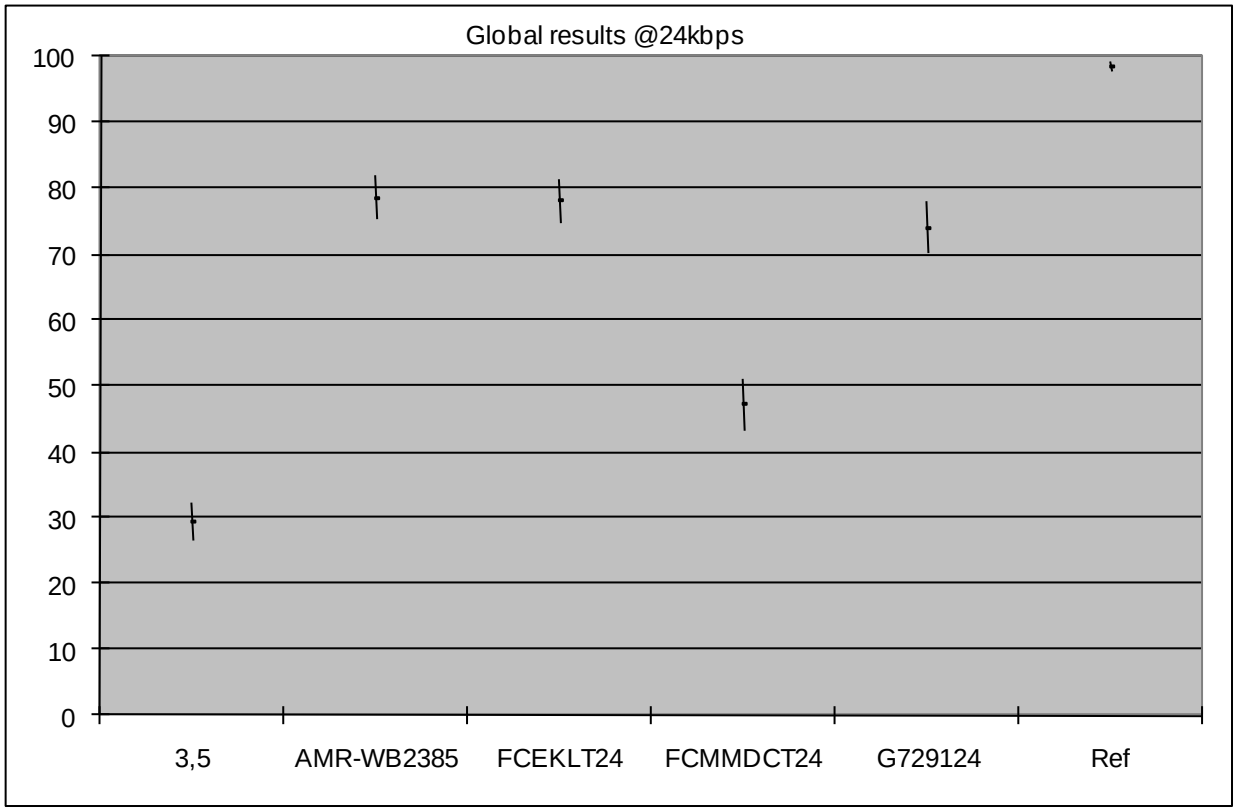


Figure 27: Results on all 12 items (24kbps – Orange Labs)

Figure 27 shows the results over the 12 audio items for all listeners.

The FCRMDCT codec is judged "fair", below that of the 3 other codecs. The FCEKLT codec is scored the same as the AMR-WB one (both "good" quality, [60-80]) and a bit higher than the G729.1 scored "good" as well.

Student	3,5	AMR-WB2385	FCEKLT24	FCMMDCT24	G729124	Ref
3,5	1,00					
AMR-WB2385	0,00	1,00				
FCEKLT24	0,00	**0,81**	1,00			
FCMMDCT24	0,00	0,00	0,00	1,00		
G729124	0,00	**0,08**	**0,13**	0,00	1,00	
Ref	0,00	0,00	0,00	0,00	0,00	1,00

Table 11: Student Test on all 12 items (24kbps – Orange Labs)

Table 11 gives the results of the Student T test.

Reminder: this test is used to observe whether the quality of a coder is significantly different from that of the other coders. A number higher than 0.05 means that the two compared configurations are not statistically different.

Consequently, figures in table 11 indicate that:

1. The FCEKLT codec at 24kbps and the AMR-WB codec at 23.85kbps are not significantly different.
2. The FCEKLT codec at 24kbps and the G729.1 codec at 24kbps are not significantly different.
3. The G729.1 codec at 24kbps and the AMR-WB codec at 23.85kbps are not significantly different.

b. Ericsson

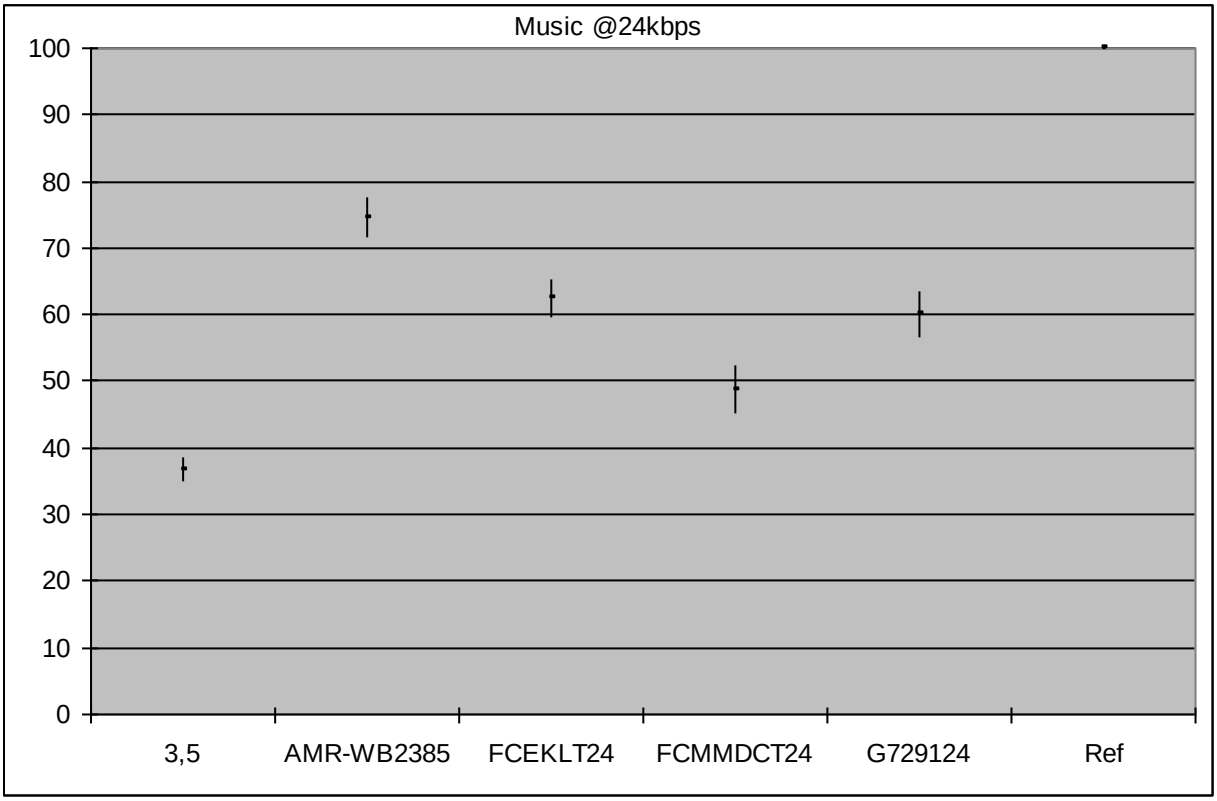


Figure 28: Results on Music items (3 items) (24kbps – Ericsson)

Figure 28 shows the results for the music items for all 7 listeners. The FCEKLT and G729.1 receive similar scores. The AMR-WB codec scores about 11 MUSHRA points higher than the FCEKLT codec. The FCRMDCT codec scores about 11 points lower then G.729.1.

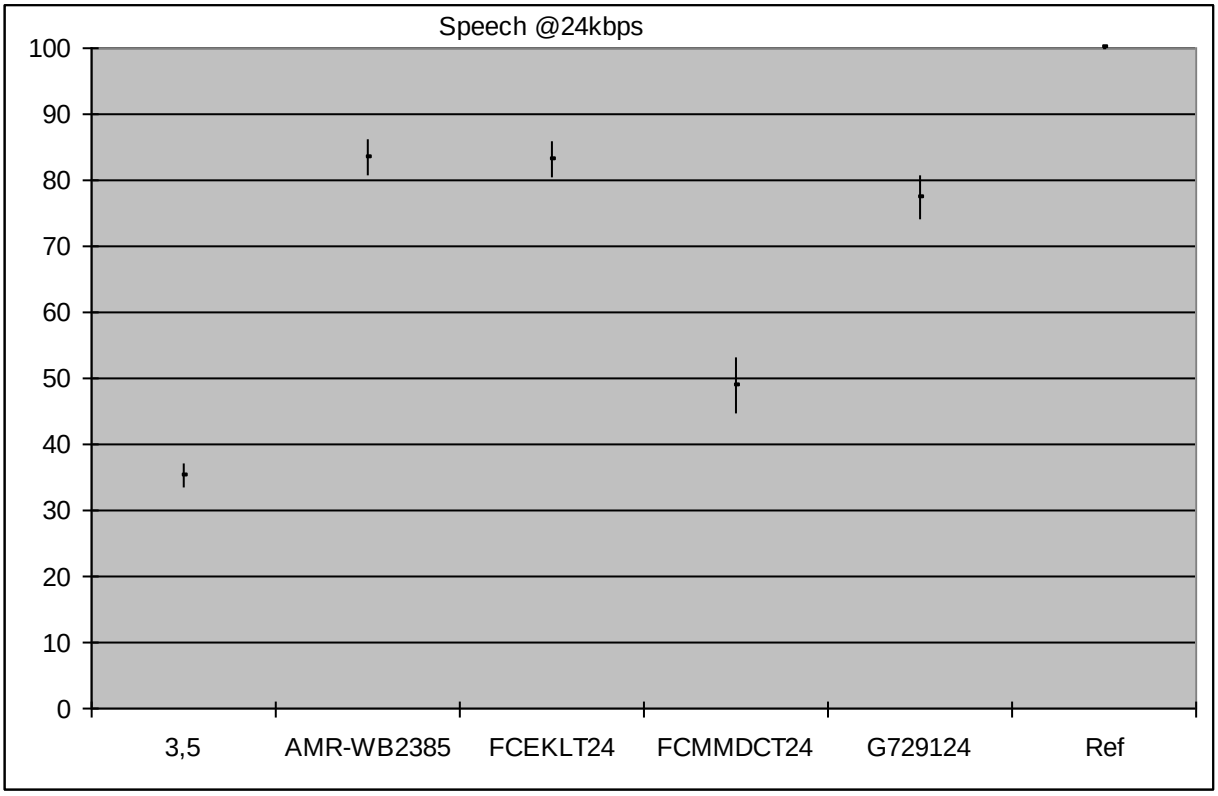


Figure 29: Results on Speech items (4 items) (24kbps – Ericsson)

The results in Figure 29 indicate on-par performance for AMR-WB and FCEKLT for speech items. Both outperform G.729.1 which outperforms the FCRMDCT codec.

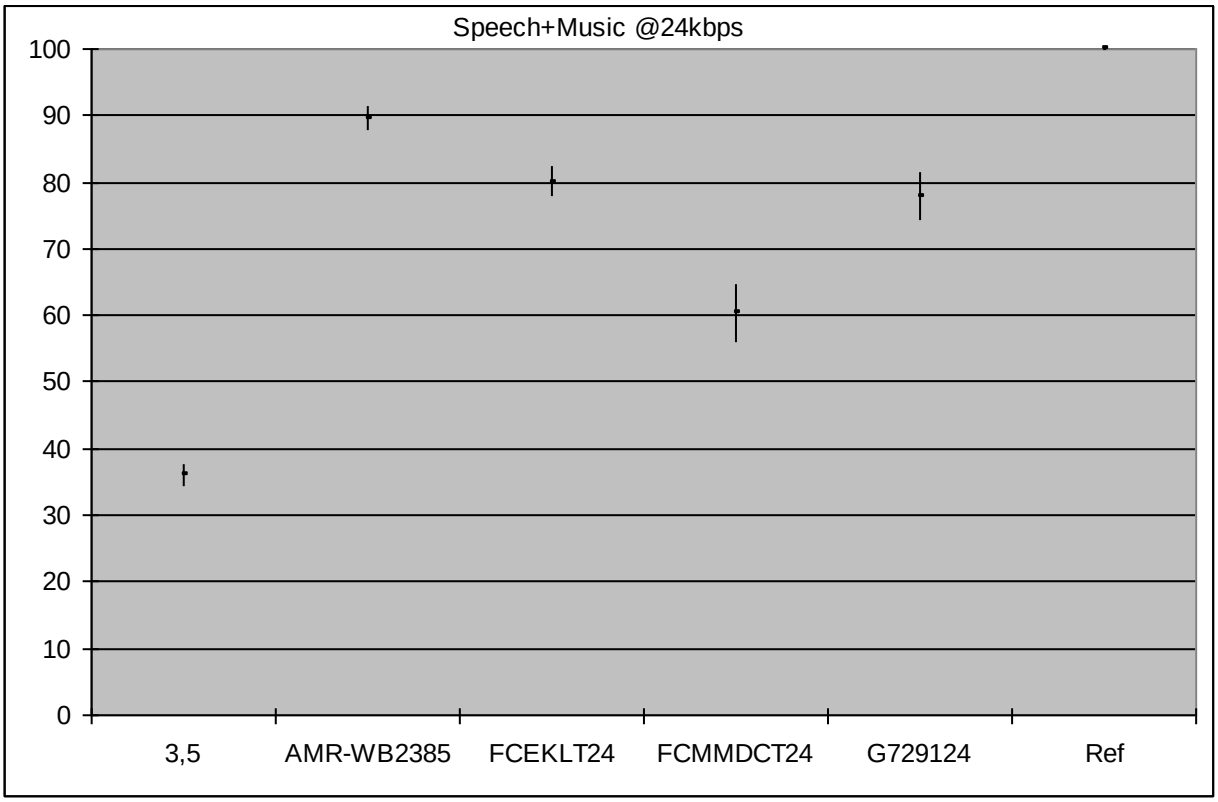


Figure 30: Results on Speech + Music items (3 items) (24kbps – Ericsson)

For speech+music items FCEKLT performs on-par with G.729.1. Both are outperformed by AMR-WB. The FCRMDCT codec is about 18 points below G.729.1

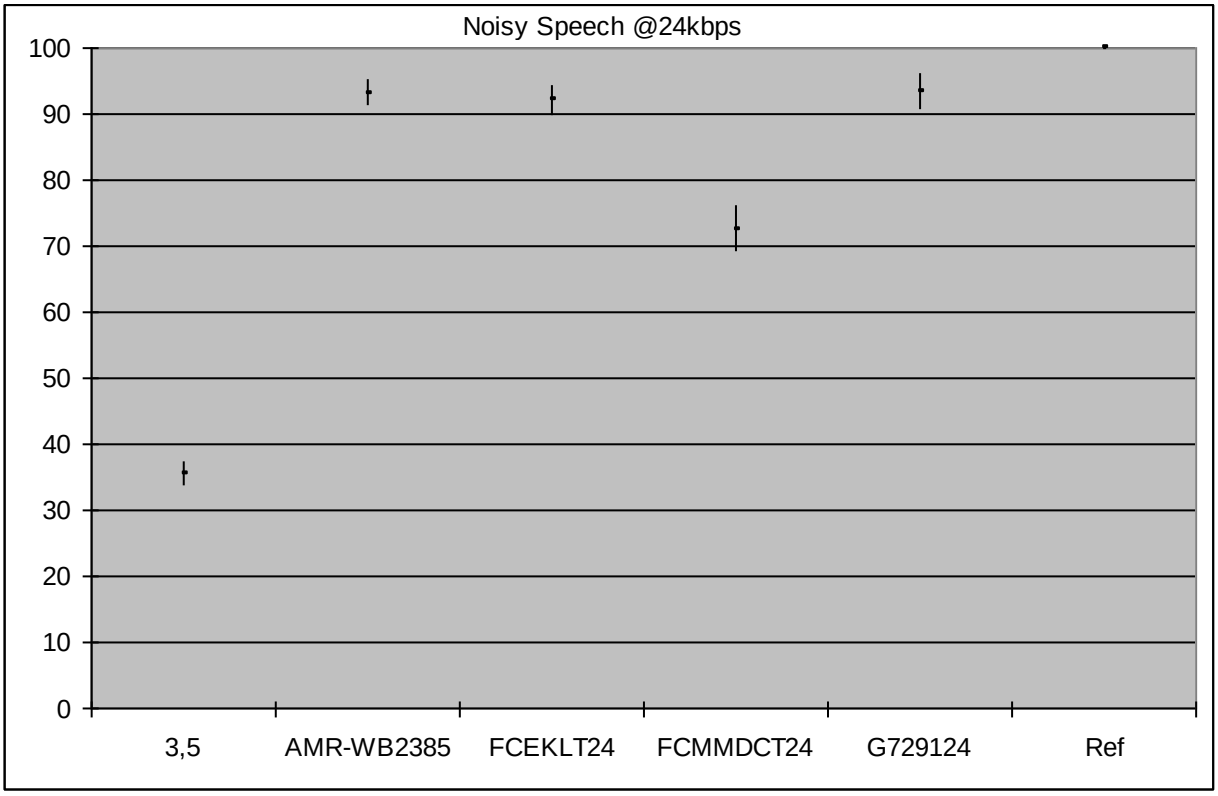


Figure 31: Results on Noisy speech items (2 items) (24kbps – Ericsson)

For noisy speech AMR-WB, G.729.1 and FCEKLT perform similar. The performance of FCRMDCT is about 20 points lower.

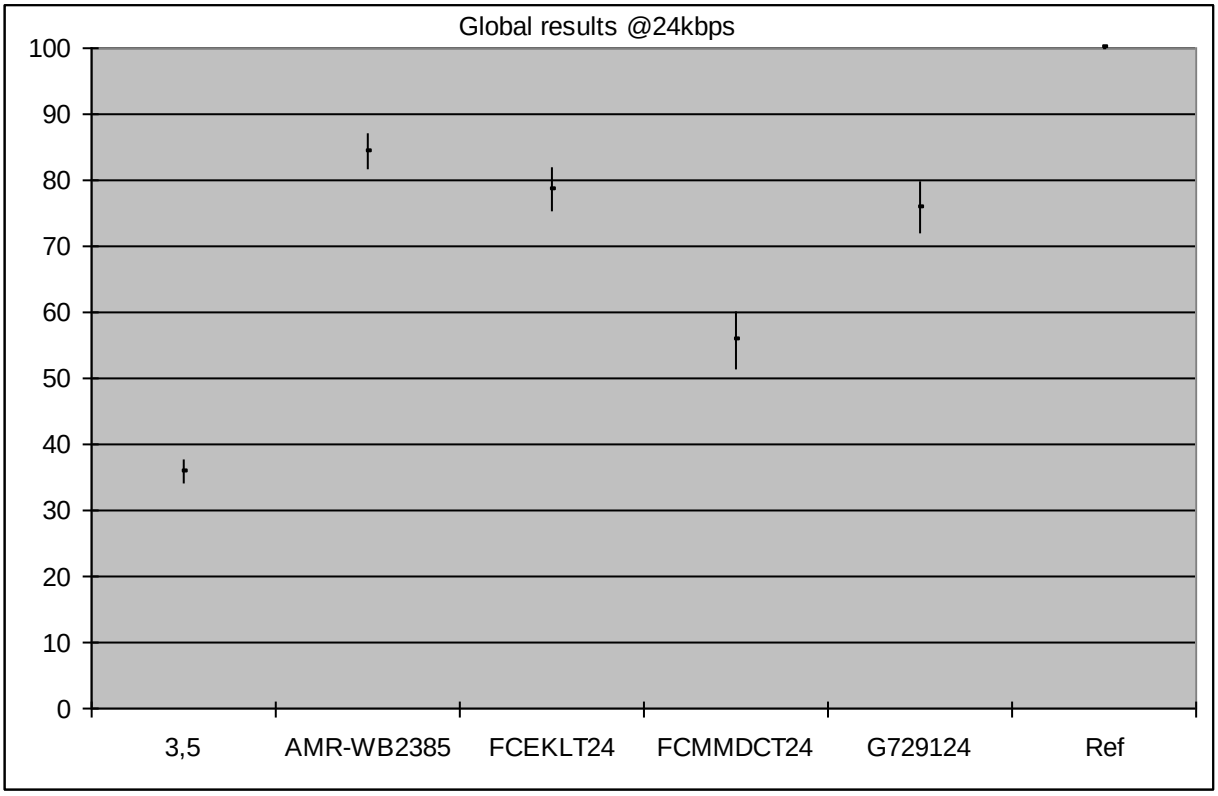


Figure 32: Results on all 12 items (24kbps – Ericsson)

Figure 32 shows a clear improvement of codec performance when increasing the rate from 14.25 kbps to 24 kbps for all codecs under test. Averaged over all items, the average score of the AMR-WB codec is higher than the average score of all the other codecs. As shown in Table 12, the difference is larger than the 95% confidence interval. The performance of the FCEKLT and G.729.1 is not separated by the 95% confidence interval.

Student	3,5	AMR-WB2385	FCEKLT24	FCMMDCT24	G729124	Ref
3,5	1,00					
AMR-WB2385	0,00	1,00				
FCEKLT24	0,00	**0,01**	1,00			
FCMMDCT24	0,00	0,00	0,00	1,00		
G729124	0,00	**0,00**	**0,29**	0,00	1,00	
Ref	0,00	0,00	0,00	0,00	0,00	1,00

Table 12: Student Test on all 12 items (24kbps – Ericsson)

c. Nokia

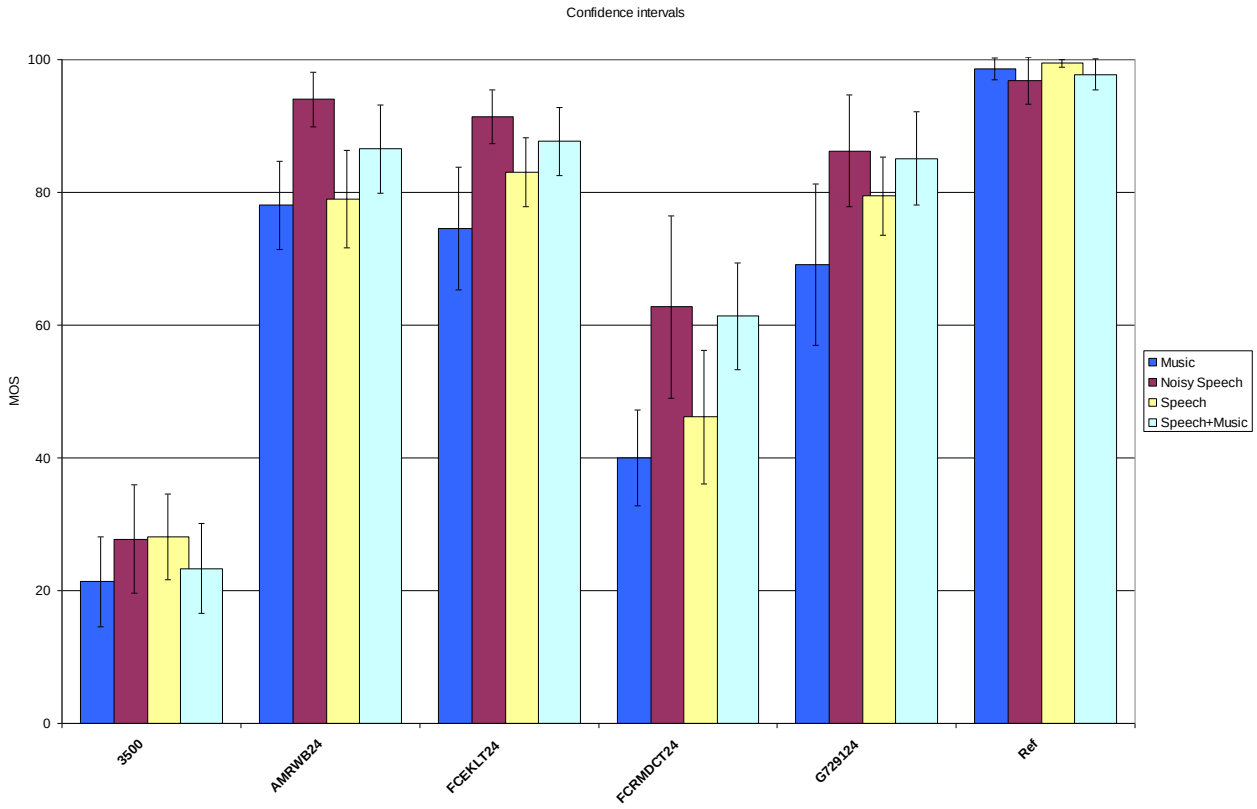


Figure 33: Results on Music items (3 items) (24kbps – Nokia)

Figure 33 shows the results for the 3 music items 4 speech items, 3 speech+music and 2 noise speech items over all 6 listeners for 24kbps. In music, the FCEKLT is better than G729.1 but does not reach AMR-WB level. In indicate that FCEKLT is slightly better than AMR-WB. For speech+music items FCEKLT performs slightly better than AMR-WB and G.729.1. For noisy speech AMR-WB and FCEKLT perform similar.

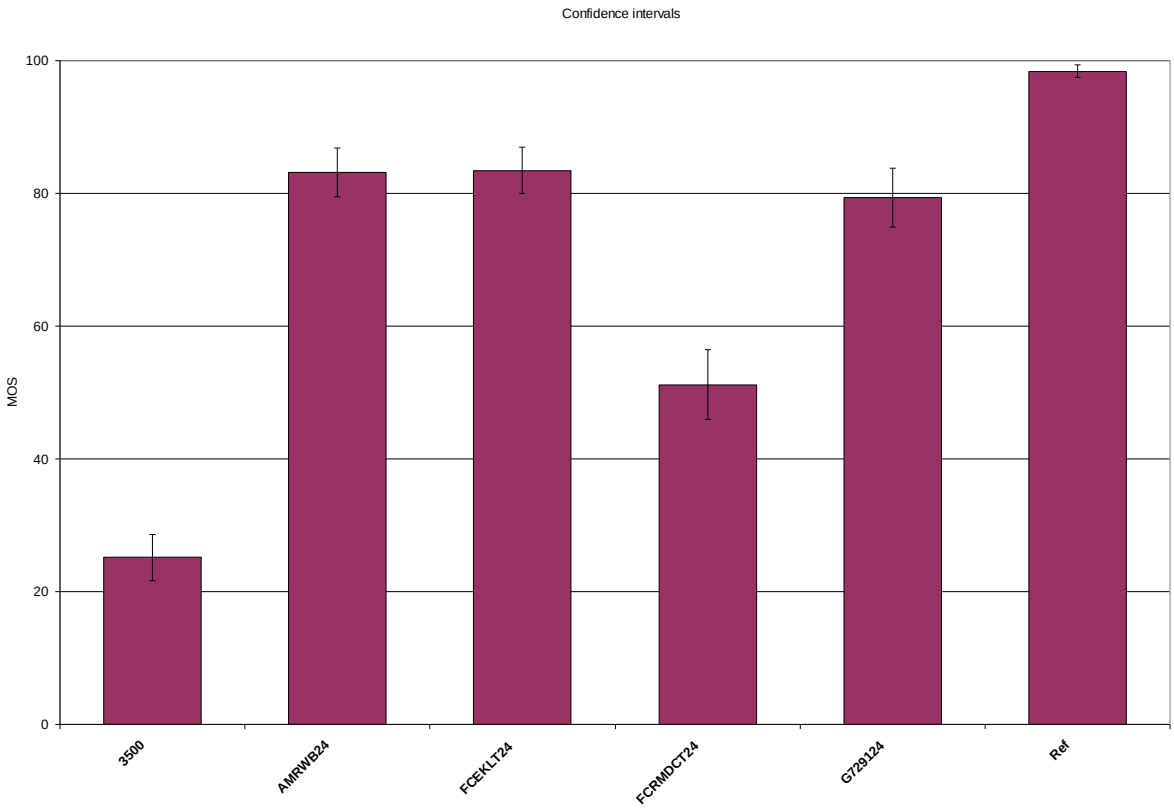


Figure 34: Results on all 12 items (24kbps – Nokia)

Figure 34 shows that at 24 kbps range FCEKLT get better score than AMR-WB. However, Table 13 indicates that the difference is within the 95% confidence interval.

	3500	*AMRWB24*	*FCEKLT24*	*FCRMDCT24*	*G729124*	*Ref*
3500						
AMRWB24	BT					
FCEKLT24	BT	EQ				
FCRMDCT24	BT	WT	WT			
G729124	BT	EQ	EQ	BT		
Ref	BT	BT	BT	BT	BT	

Table 13: Student Test on all 12 items (24kbps – Nokia)

3. High bit rate: 32kbps

a. Orange Labs

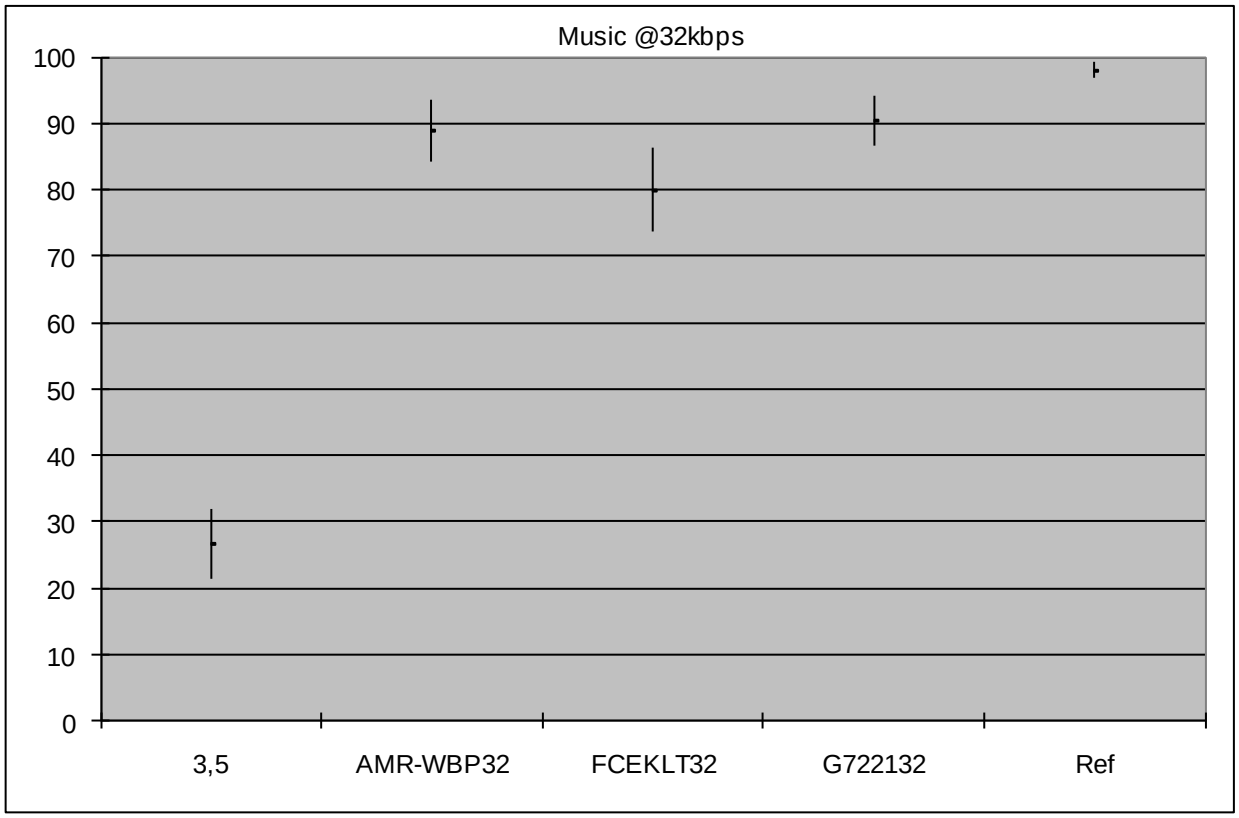


Figure 35: Results on Music items (3 items) (32kbps – Orange Labs)

Figure 35 shows the results over the 3 music items for all listeners at 32kbps. The FCEKLT has been scored on the border between "good" (60-80) and "excellent" (80-100) just below the G722.1 and the AMR-WB both judged "excellent" (80-100).

Remark: The reference items have been scored on the average a bit lower than 100; that shows the overall excellent quality of all the tested codecs as sometimes listeners couldn't distinguish between the reference items and the coded ones.

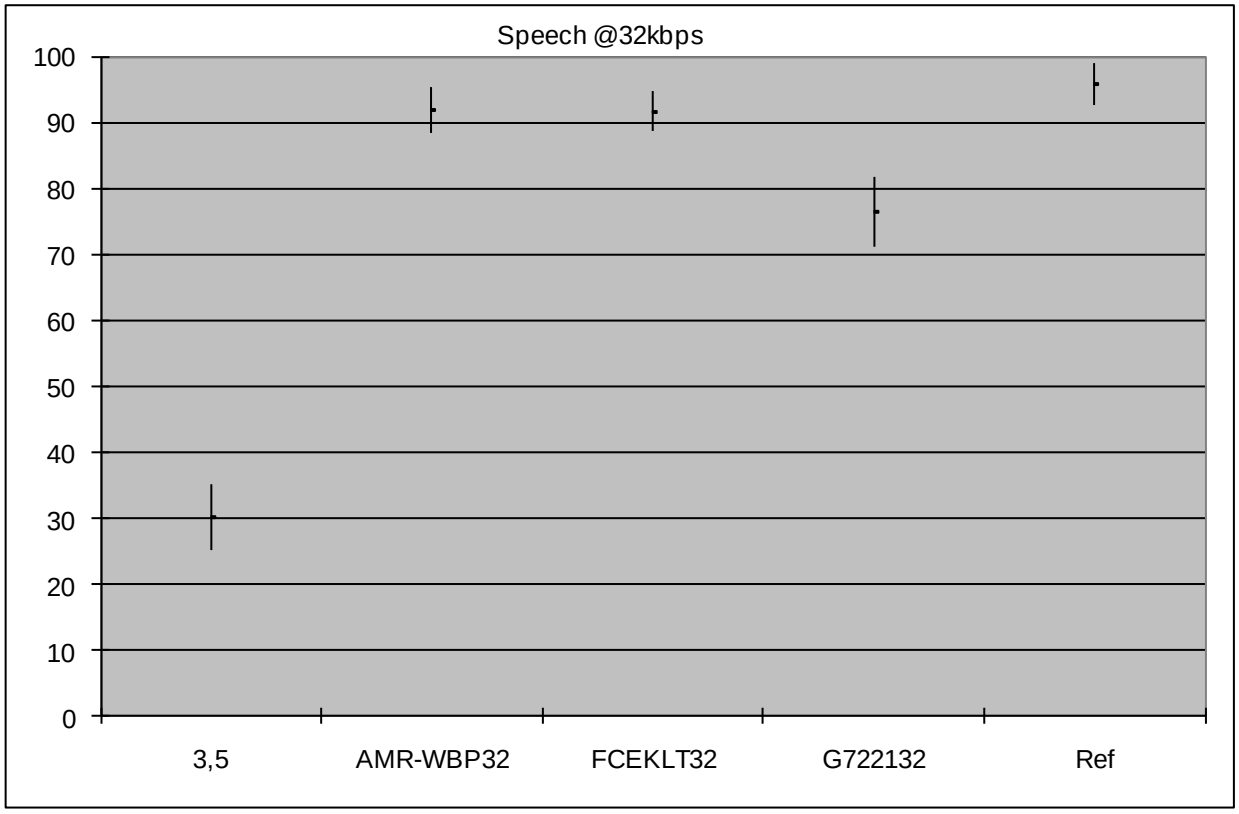


Figure 36: Results on Speech items (4 items) (32kbps – Orange Labs)

Figure 36 shows the results at 32kbps for the 4 speech items over all the listeners. The quality of the FCEKLT and AMR-WB codecs have been judged the same that is "excellent" (80-100), while that of the G722.1 codec is assessed as "good" (60-80).

The same remark as above applies about the reference items.

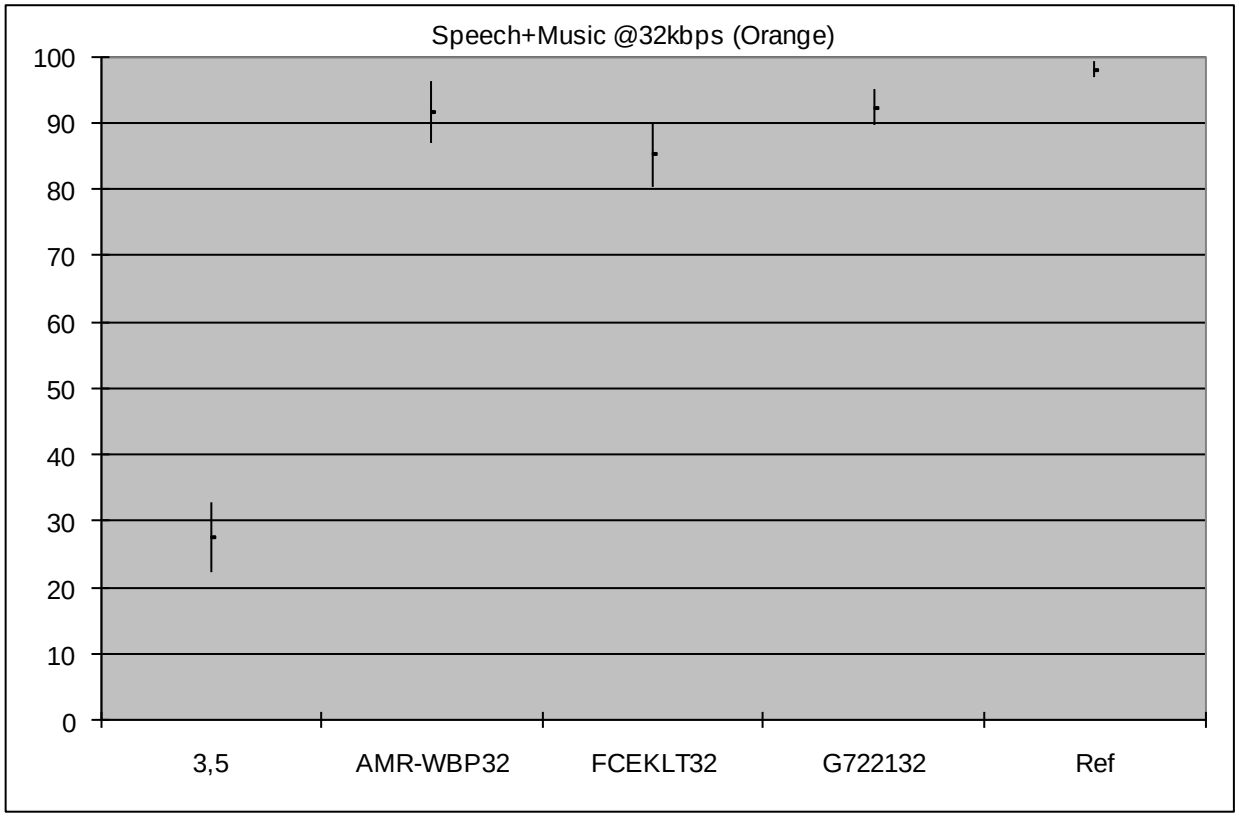


Figure 37: Results on Speech + Music items (3items) (32kbps – Orange Labs)

Figure 37 shows the results at 32kbps for the 3 speech and music items over all listeners. All 3 tested codecs have been scored "excellent" (80-100), the FCEKLT codec being assessed a bit lower than the 2 others.

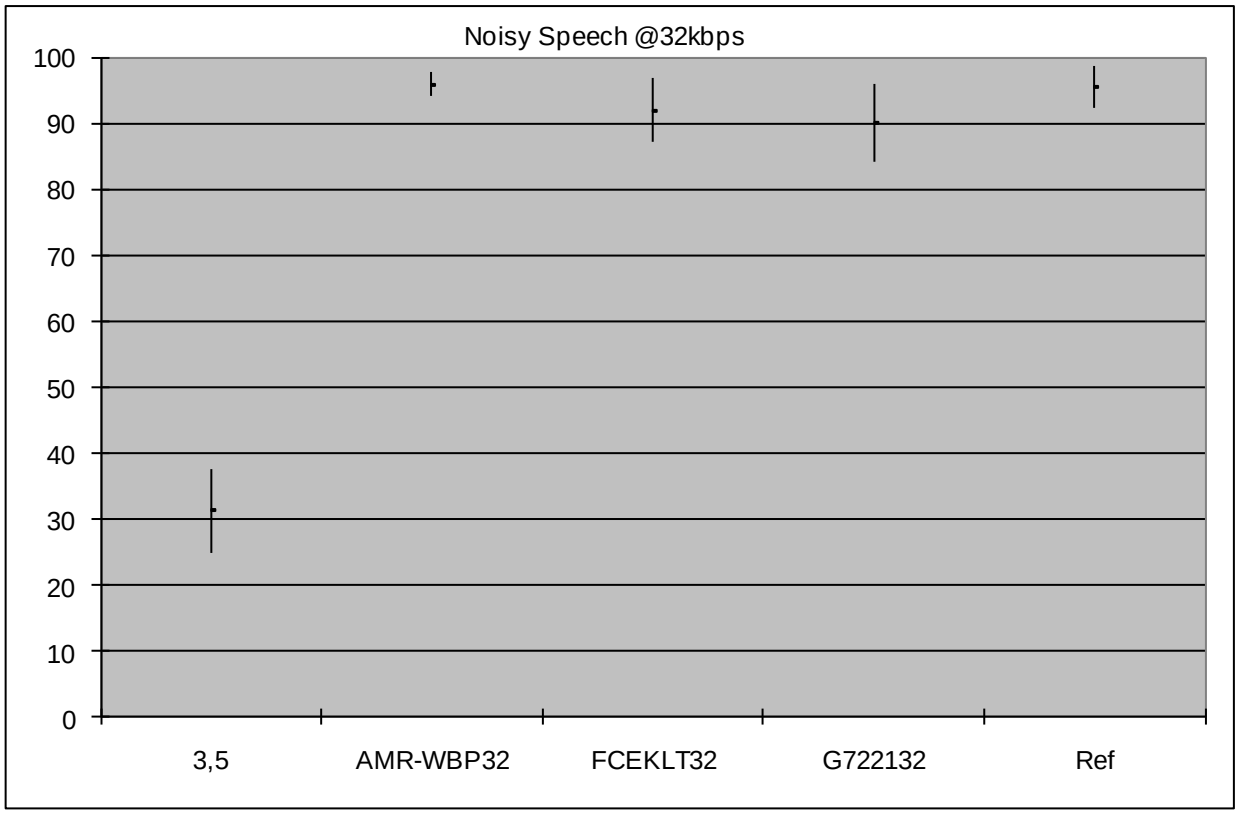


Figure 38: Results on Noisy speech items (2 items) (32kbps – Orange Labs)

Figure 38 shows the results at 32kbps for the 2 noisy speech items over all the listeners. The qualities of the FCEKLT and the AMR-WB codecs have been judged nearly the same that is “excellent” (80-100), as for the G722.1 codec which is a bit lower.

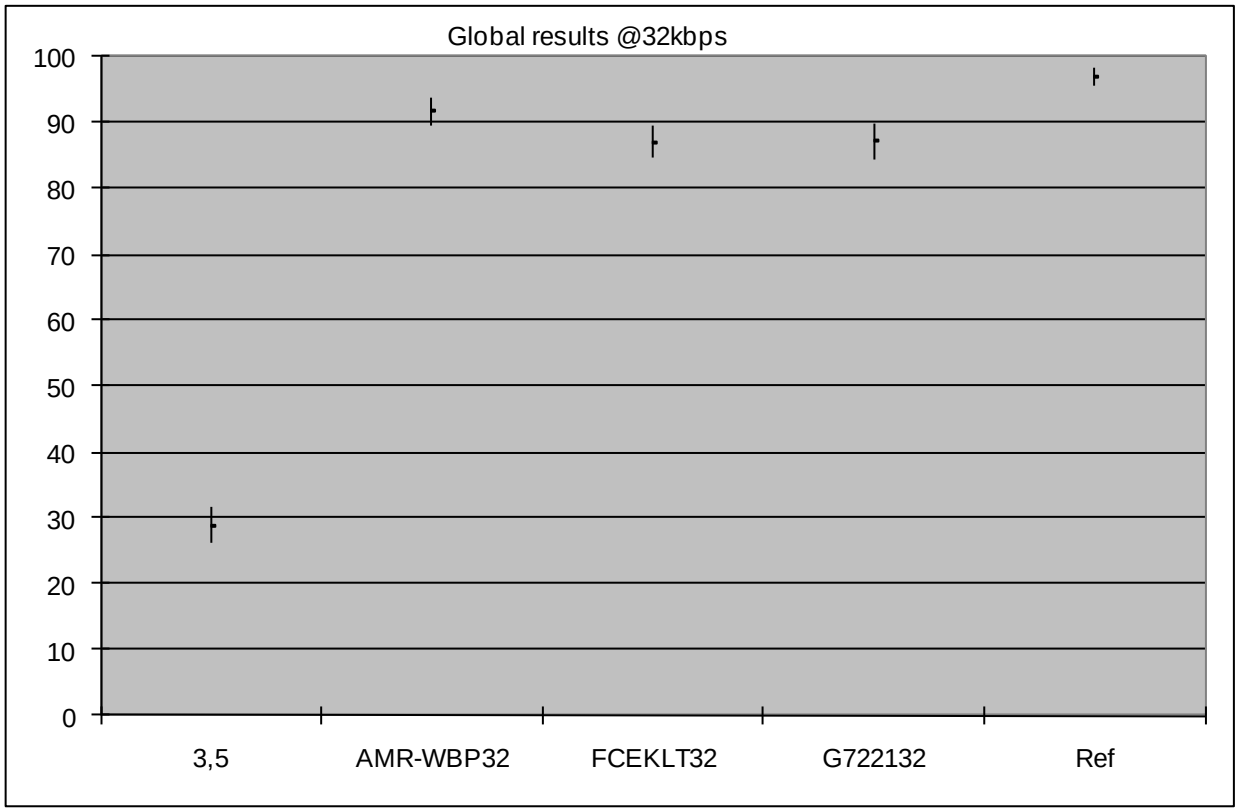


Figure 39: Results on all 12 items (32kbps – Orange Labs)

Figure 39 shows the results over the 12 audio items for all listeners.

The quality of all tested codecs is judged “excellent” (80-100), the one of the AMR-WB codec being a bit higher. The same remark as above can be made about reference items quality.

Student	3,5	AMR-WB32	FCEKLT32	G722132	Ref
3,5	1,00				
AMR-WBP32	0,00	1,00			
FCEKLT32	0,00	0,01	1,00		
G722132	0,00	0,00	**0,62**	1,00	
Ref	0,00	0,00	0,00	0,00	1,00

Table 14: Student Test on all 12 items (32kbps – Orange Labs)

Table 14 gives the results of the Student T test.

Reminder: this test is used to observe whether the quality of a coder is significantly different from that of the other coders. A number higher than 0.05 means that the two compared configurations are not statistically different.

Consequently, figures in table 14 indicate that the FCEKLT codec at 32kbps and the G722.1 codec at 32kbps are not significantly different.

b. Ericsson

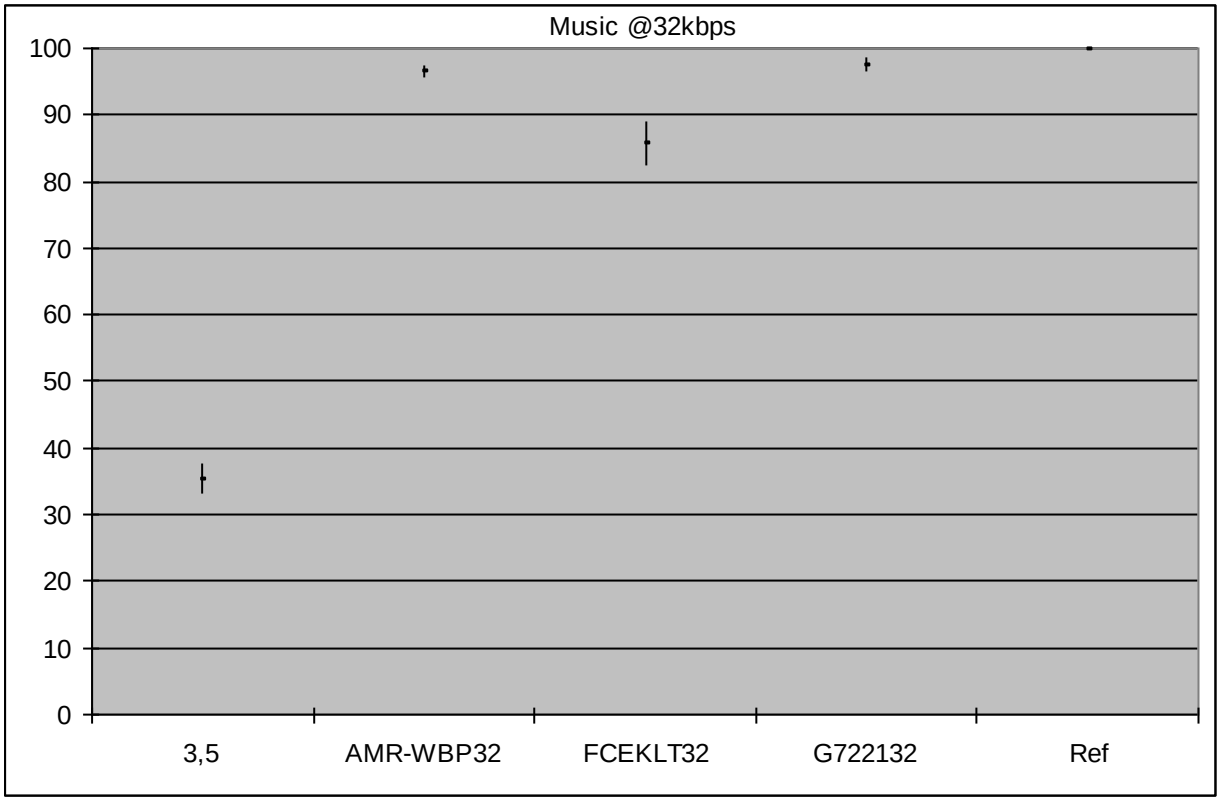


Figure 40: Results on Music items (3 items) (32kbps – Ericsson)

Figure 40 shows the results for the music items for all 6 listeners at a rate of 32 kbps. The two reference codecs perform equally well and are scored about 10 points higher than the FCEKLT codec. The hidden reference was not recognized by all listeners. However, its score was always close to 100.

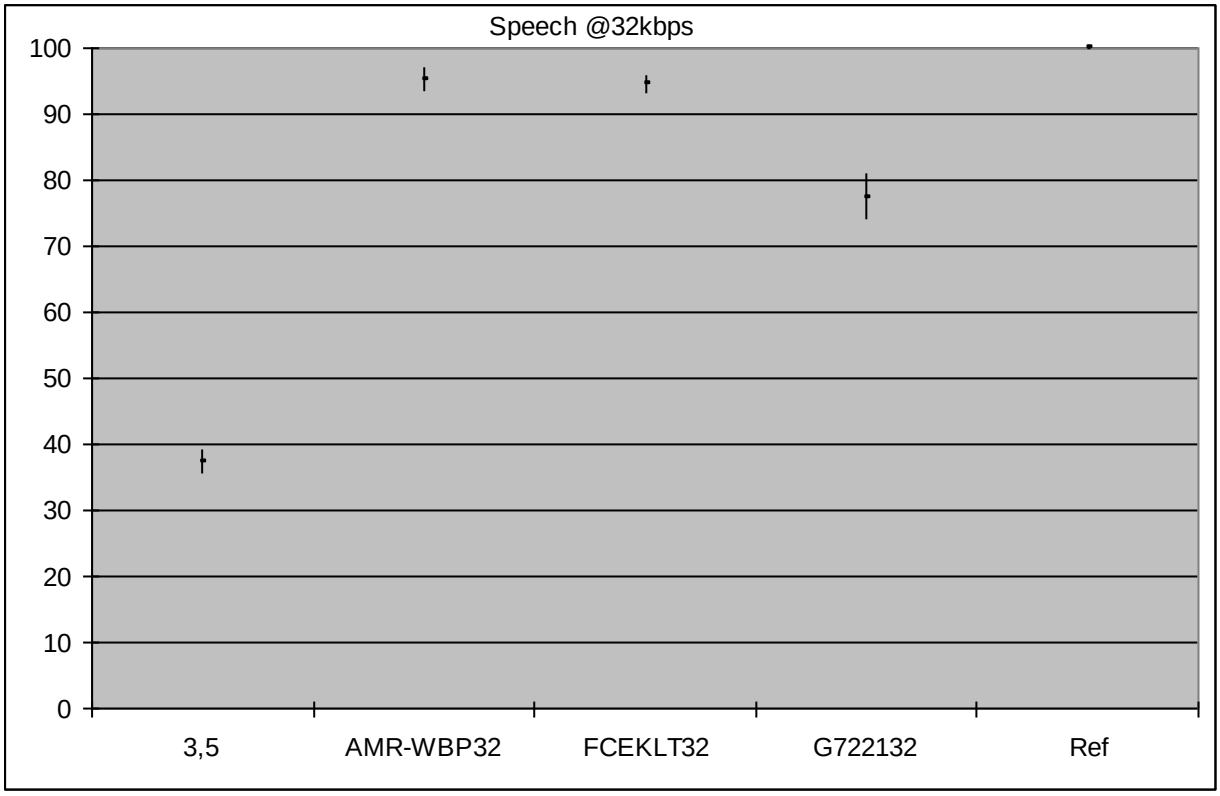


Figure 41: Results on Speech items (4 items) (32kbps – Ericsson)

For speech items the AMR-WB+ and FCEKLT perform equally well. The G.722.1 codec scores about 17 points lower.

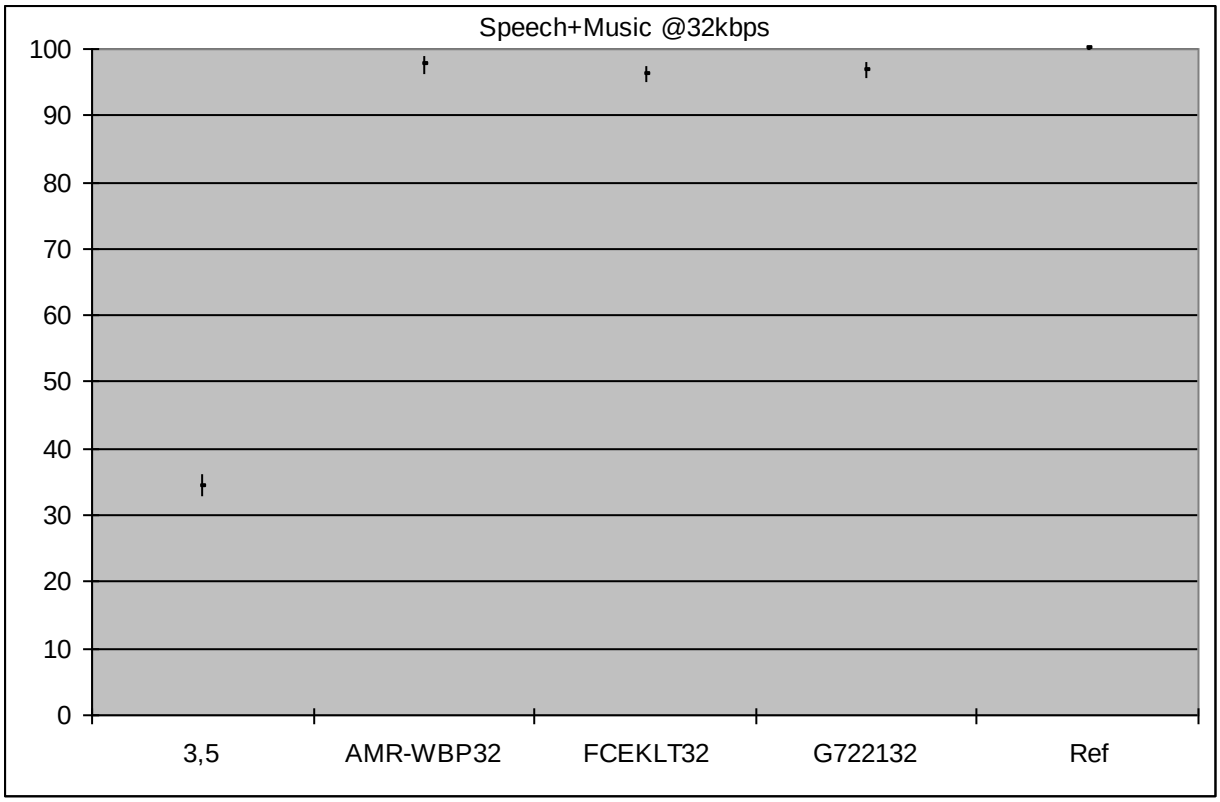


Figure 42 Results on Speech + Music items (3items) (32kbps – Ericsson)

For speech+music items the performance of all three codecs was very close.

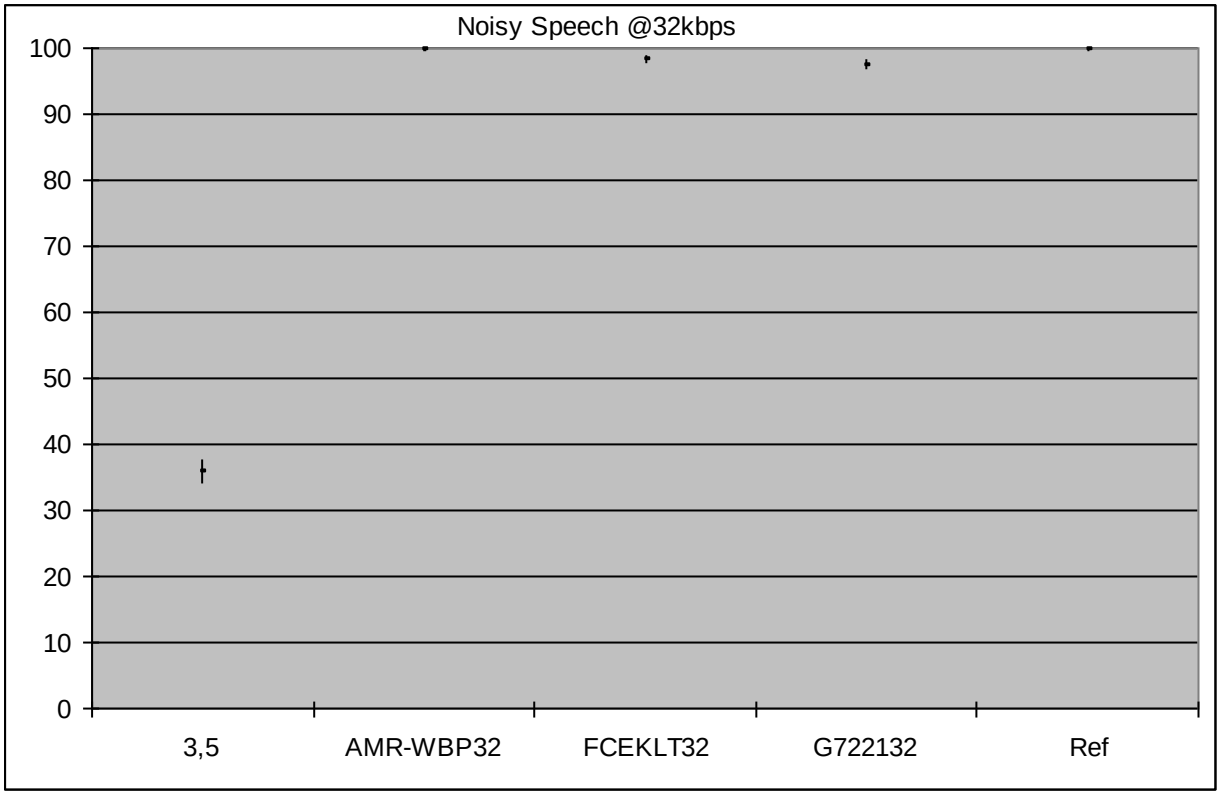


Figure 43: Results on Noisy speech items (2 items) (32kbps – Ericsson)

The same was observed for noisy speech items.

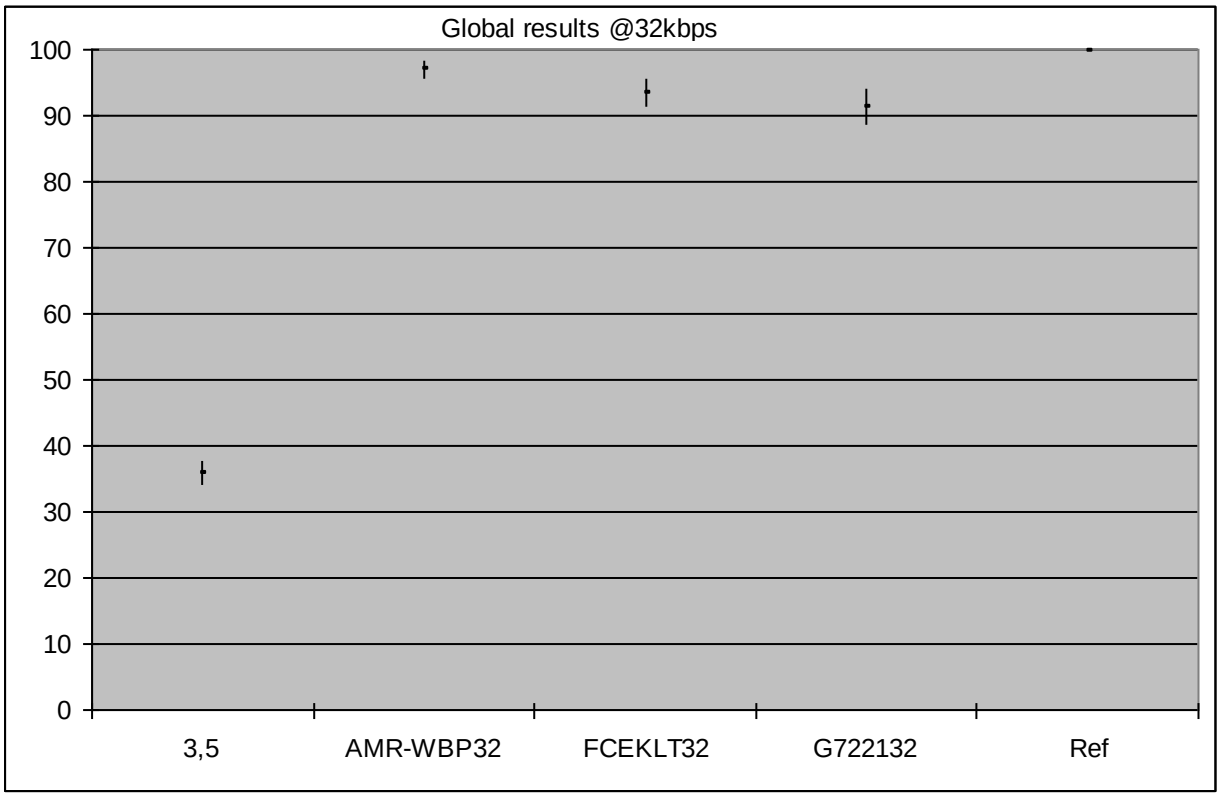


Figure 44: Results on all 12 items (32kbps – Ericsson)

The average scores for all items and all 6 listeners indicate a small performance advantage of AMR-WB+, this is confirmed by the Student test in Table 15. The FCE-KLT performance can not be separated from the G.722.1 performance as Table 15 shows.

Student	3,5	AMR-WB32	FCEKLT32	G722132	Ref
3,5	1,00				
AMR-WBP32	0,00	1,00			
FCEKLT32	0,00	0,02	1,00		
G722132	0,00	0,00	**0,21**	1,00	
Ref	0,00	0,00	0,00	0,00	1,00

Table 15: Student Test on all 12 items (32kbps – Ericsson)

c. Nokia

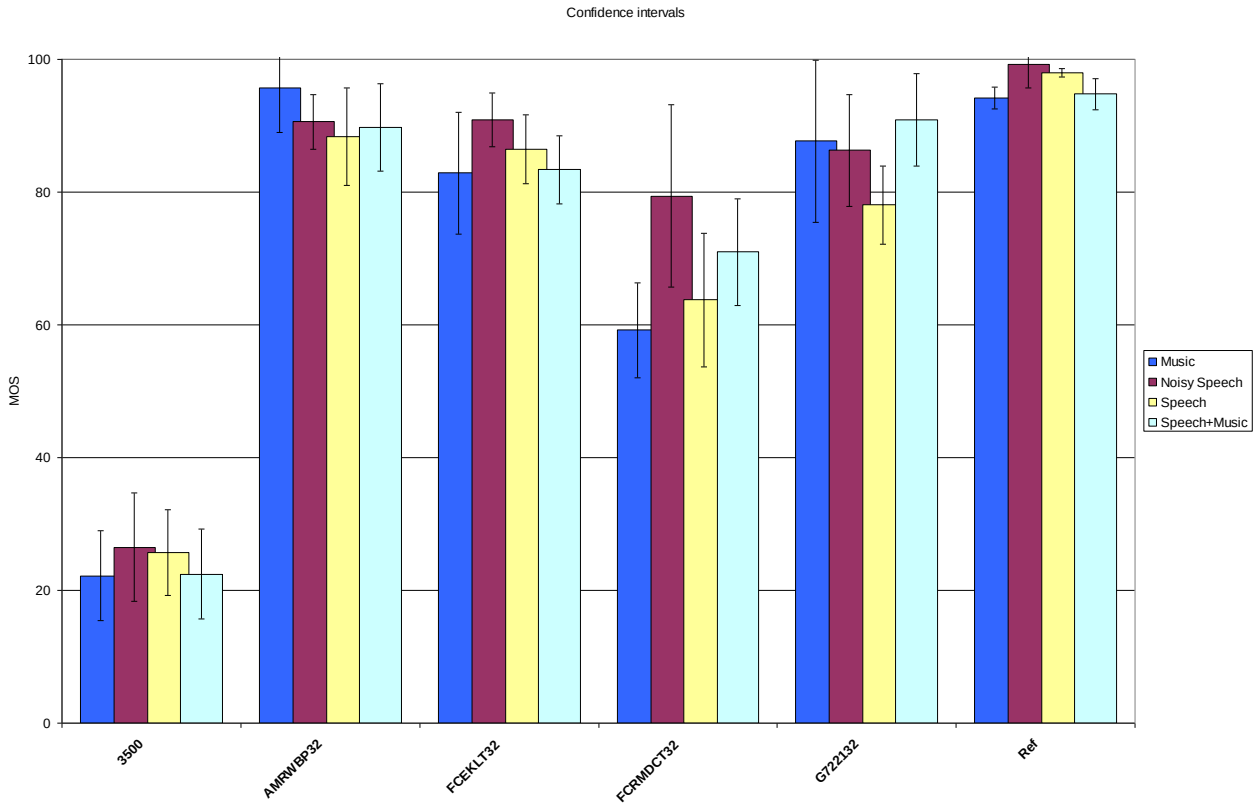


Figure 45: Results on Music items (3 items) (32kbps – Nokia)

Figure 45 shows the results for the 3 music items 4 speech items, 3 speech+music and 2 noise speech items over all 5 listeners for 32 kbps. In music, AMR-WB+ scored best and about 13 points higher than FCEKLT codec. For items the AMR-WB+ and FCEKLT perform about equally well. The overall performance of all codecs is so good that some of the listeners had difficulties finding the hidden reference. For speech+music items the performance the reference codecs is very equal. FCEKLT is about 5 points below. AMR-WB+ and FCEKLT are almost equal in noisy speech items.

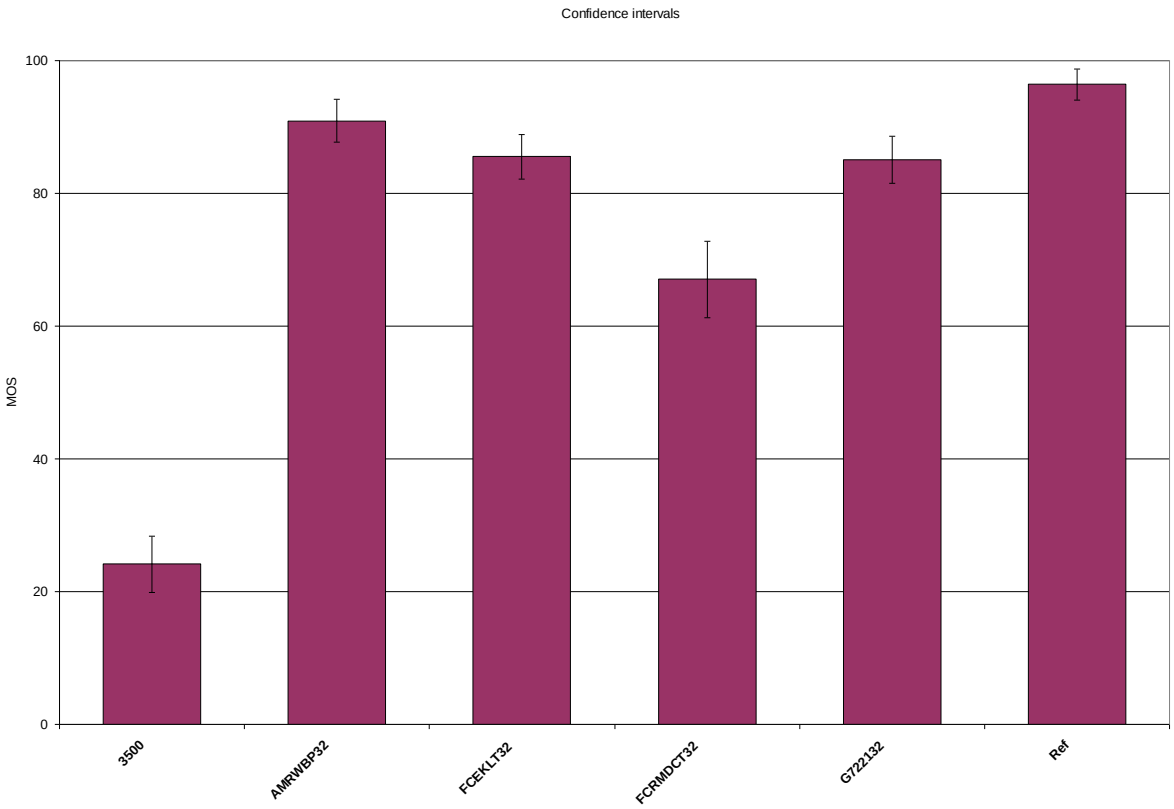


Figure 46: Results on all 12 items (32kbps – Nokia)

The average scores for all items and all 5 listeners indicate a small performance advantage of AMR-WB+. However, the Student test in Table 16. The FCE-KLT performance can not be separated from the G.722.1 performance as Table 16 shows.

	3500	*AMRWBP32*	*FCEKLT32*	*FCRMDCT32*	*G722132*	*Ref*
3500						
AMRWBP32	BT					
FCEKLT32	BT	EQ				
FCRMDCT32	BT	WT	WT			
G722132	BT	WT	EQ	BT		
Ref	BT	BT	BT	BT	BT	

Table 16: Student Test on all 12 items (32kbps – Nokia)

11.Global "Quality" Results

The first remark to be made is that the overall quality of the *FlexCode* codec with constrained entropy and the KLT transform (FCEKLT) has increased significantly since the intermediate test [1] a year ago.

Comparing both *FlexCode* codecs, FCEKLT and FCRMDCT, it is seen that the latter performs worse than the FCEKLT codec. This is expected given that the focus in WP1 development was on the KLT codec. Consequently, we concentrate on the FCEKLT results also in this report. We note that a FCRKLT (*FlexCode* constrained-resolution KLT based) coder exists, but was not tested because limitations in testing slots.

All labs (Orange, Ericsson and Nokia) report the same trends. However, the scores reported from Ericsson are approximately 10 points lower than the ones reported by the Orange and Nokia labs. The reason for this difference is not clear. We observe that for the 32kbps samples the Ericsson listeners consistently identified the hidden reference while scoring the coded signals lower than 100 marks while confusion of hidden reference and coded signals happened more frequently at Orange labs. Thus, it is unlikely that the Ericsson listeners were simply in-sensitive to the distortions at hand. At Nokia test site, the experiments were conducted by naïve listeners. This partly explains the scoring of reference signal below 100 in 32 kbps experiment. Due to a processing error, FCRMDCT version was also tested in the 32 kbps test. This should not, however, affect the overall results which are in line with Orange and Ericsson test results.

For low bit rate (14.25kbps), the FCEKLT is judged in the same quality range (on all items) as the "state of the art codec" AMR-WB, that is "fair" (40-60), although scored 8 marks lower at Orange labs. At Ericsson the difference was of 7 marks and both codecs reached "good" performance. At Nokia the difference was around 6 marks and both codecs reached "good" performance as well. The music items are the weakness of the FCEKLT at such a bit rate, but the same can be note for the AMR-WB.

At medium bitrates (24kbps), globally, the FCEKLT is not distinguishable from the state of the art codec G729.1. Comparing the FCEKLT codec to AMR-WB the results differ slightly between the three test sites. At Orange labs and Nokia the two codecs were not distinguishable while at Ericsson a 6 MUSHRA point difference to the advantage of AMR-WB makes them distinguishable. Furthermore, one can note that on pure speech items both Orange labs and Nokia report an advantage of FCEKLT over both reference codecs, while Ericsson reports on-part performance with AMR-WB and a 5 point advantage over G.729.1

At high bitrates (32kbps), globally, the FCEKLT is not distinguishable from G.722.1. Comparing to AMR-WB+, Nokia report no statistical differences while the test results from Ericsson and Orange labs show a statistical significant difference of 3 to 6 MUSHRA points to the advantage of AMR-WB+. Nevertheless, for music all sites report a 10 MUSHRA point difference between the two reference codec and FCEKLT to the advantage of the reference codecs.

Detailed scores figures can be found in annexes 1 through 4.

V. CONCLUSION

The test of the channel coder shows that the *FlexCode* codec with constraint entropy and the KLT transform behaves slightly better than the G729.1 when facing packet losses. In parallel, it has been shown that the bit error adaptation allows a better quality with the *FlexCode* codec with constrained resolution and the MDCT transform.

The quality of the *FlexCode* source codecs has improved since the previous quality tests. This is especially the case for the *FlexCode* codec with constrained entropy and the KLT transform. The FCEKLT is now comparable to AMR-WB and AMR-WB+ at all bit-rates within the tested range; it is comparable as well to G729.1 at 24kbps and to G722.1 at 32kbps.

The *FlexCode* source codecs and the reference codecs (AMRWB, G.729.1, G.722.1, AMR-WB+) were tested at equivalent bit rates (14, 24 and 32 kbit/s). However, it is worth noting that these codecs have different coding attributes, such as coding delay and frame size, and that the tradeoffs between all coding attributes have an influence on the achievable quality. In particular, we note that the AMRWB+ coder has a significantly longer delay than the tested FCEKLT coder.

We also note that we have made an improvement to the distribution-preserving quantizer (cf. D1.2) after the test. The distribution-preserving quantizer should be used only for signal components where the distribution (signal model) is accurate and this means it should not be used for the spectral fine structure. The change likely would result in an improved score for the FCEKLT coder at low rates.

Informal tests not reported here indicate that the difference in subjective performance of FCEKLT and the constrained-resolution based KLT coder is small.

Altogether, the *FlexCode* codec with constrained entropy and the KLT transform is now comparable to the state-of-the-art codecs over the range of tested bit rates. It is also comparable in terms of countering channels errors and packet losses. In contrast to existing coders the *FlexCode* codec can be redesigned in real-time to provide optimal performance at all times.

VI. REFERENCES

[1]: WP5: Intermediate Subjective Quality Tests – report D5.1

[2]: WP5: Quality Assessment plan – Final Test

[3]: WP5: Processing test plan – Phase 1

[4]: WP1: Baseline Source Coder – report D1.1

[5]: ITU-T G.729.1 "G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bit stream interoperable with G.729"

[5bis]: ITU-T G.729.1 Amendment 3 "Extension of the G.729.1 low delay mode functionality to 14 kbit/s, and corrections to the main body and Annex B"

[6]: WP2: Final Channel Coder – report D2.3

[7]: ITU-T Q12/7 recommandation P.880, May 2004

[8]: ITU-R Recommendation BS.1534 (June 2001)/ Method for the subjective assessment of intermediate quality level of coding systems

[9]: WP4: Report on *FlexCode* Test Material Processing – report D4.1

Annex 1: test results arrays - Orange Labs test site

1. Low bit rate: 14kbps (Orange Labs)

Music	Av/codecs/it	CI95
3,5	33,88	5,58
AMR-WB1425	33,36	5,24
FCEKLT14	23,39	3,84
FCMMDCT14	15,15	4,14
Ref	100,00	0,00

Table 1: Results on Music items (14.25kbps – Orange Labs)

Speech	Av/codecs/it	CI95
3,5	36,70	5,18
AMR-WB1425	67,50	6,10
FCEKLT14	61,39	6,27
FCMMDCT14	22,05	4,23
Ref	100,00	0,00

Table 2: Results on Speech items (14.25kbps – Orange Labs)

Speech+Music	Av/codecs/it	CI95
3,5	31,52	5,01
AMR-WB1425	53,55	6,61
FCEKLT14	51,97	6,73
FCMMDCT14	30,94	5,45
Ref	100,00	0,00

Table 3: Results on Speech + Music items (14.25kbps – Orange Labs)

Noisy Speech	Av/codecs/it	CI95
3,5	33,82	6,44
AMR-WB1425	76,73	7,51
FCEKLT14	65,05	8,66
FCMMDCT14	34,59	8,27
Ref	99,59	0,59

Table 4: Results on Noisy speech items (14.25kbps – Orange Labs)

	Average/cod	CI95
3,5	34,22	2,76
AMR-WB1425	57,02	4,13
FCEKLT1425	50,14	4,19
FCMMDCT1425	24,64	2,85
Ref	99,93	0,10

Table 5: Global Results on all items (14.25kbps – Orange Labs)

2. Medium bit rate: 24kbps (Orange Labs)

Music	Av/codecs/it	CI95
3,5	26,37	6,42
AMR-WB2385	66,37	7,21
FCEKLT24	64,17	7,75
FCMMDCT24	38,47	8,07
G729124	57,07	8,35
Ref	98,90	1,24

Table 6: Results on Music items (24kbps – Orange Labs)

Speech	Av/codecs/it	CI95
3,5	31,40	5,24
AMR-WB2385	79,40	5,34
FCEKLT24	82,13	4,58
FCMMDCT24	40,30	6,09
G729124	81,68	5,35
Ref	98,73	0,86

Table 7: Results on Speech items (24kbps – Orange Labs)

Speech+Music	Av/codecs/it	CI95
3,5	26,93	4,96
AMR-WB2385	83,37	4,93
FCEKLT24	79,20	6,59
FCMMDCT24	53,33	6,38
G729124	73,13	7,86
Ref	98,33	1,74

Table 8: Results on Speech + Music items (24kbps – Orange Labs)

Noisy Speech	Av/codecs/it	CI95
3,5	34,00	7,19
AMR-WB2385	88,45	5,88
FCEKLT24	89,15	4,48
FCMMDCT24	65,10	9,00
G729124	85,70	6,25
Ref	97,25	2,33

Table 9: Results on Noisy speech items (24kbps – Orange Labs)

	Average/cod	CI95
3,5	29,46	2,94
AMR-WB2385	78,64	3,26
FCEKLT24	78,08	3,40
FCMMDCT24	47,23	3,98
G729124	74,06	3,98
Ref	98,43	0,72

Table 10: Global Results on all items (24kbps – Orange Labs)

3. High bit rate: 32kbps (Orange Labs)

Music	Av/codecs/it	S	CI95
3,5	26,67		5,18
AMR-WBP32	88,83		4,71
FCEKLT32	79,93		6,31
G722132	90,43		3,87
Ref	98,13		1,23

Table 11: Results on Music items (32kbps – Orange Labs)

Speech	Av/codecs/it	S	CI95
3,5	30,18		5,03
AMR-WBP32	91,88		3,48
FCEKLT32	91,58		3,02
G722132	76,48		5,31
Ref	95,83		3,23

Table 12: Results on Speech items (32kbps – Orange Labs)

Speech+Mus	Av/codecs/it	S	CI95
3,5	27,63		5,23
AMR-WBP32	91,57		4,61
FCEKLT32	85,23		4,88
G722132	92,40		2,76
Ref	98,03		1,22

Table 13: Results on Speech + Music items (32kbps – Orange Labs)

Noisy Speec	Av/codecs/it	S	CI95
3,5	31,30		6,36
AMR-WBP32	95,85		1,76
FCEKLT32	91,90		4,81
G722132	90,05		5,90
Ref	95,50		3,18

Table 14: Results on Noisy speech items (32kbps – Orange Labs)

	Average/cod	S	CI95
3,5	28,85		2,72
AMR-WBP32	91,70		2,08
FCEKLT32	87,13		2,53
G722132	87,18		2,64
Ref	96,90		1,28

Table 15: Global Results on all items (32kbps – Orange Labs)

Annex 2: test results arrays - Ericsson test site

1. Low bit rate: 14kbps (Ericsson)

Music	Av/codecs/ite	CI95
3,5	39,39	2,18
AMR-WB1425	54,28	2,37
FCEKLT14	49,22	1,68
FCMMDCT14	32,33	2,91
Ref	100,00	0,00

Table 1: Results on Music items (14.25kbps – Ericsson)

Speech	Av/codecs/ite	CI95
3,5	39,13	2,41
AMR-WB1425	75,21	2,27
FCEKLT14	67,75	2,78
FCMMDCT14	33,17	3,54
Ref	100,00	0,00

Table 2: Results on Speech items (14.25kbps – Ericsson)

Speech+Music	Av/codecs/ite	CI95
3,5	37,72	2,48
AMR-WB1425	70,28	2,10
FCEKLT14	62,11	2,22
FCMMDCT14	47,17	3,13
Ref	100,00	0,00

Table 3: Results on Speech + Music items (14.25kbps – Ericsson)

Noisy Speech	Av/codecs/ite	CI95
3,5	38,17	2,49
AMR-WB1425	83,17	1,74
FCEKLT14	78,25	1,96
FCMMDCT14	49,92	3,25
Ref	100,00	0,00

Table 4: Results on Noisy speech items (14.25kbps – Ericsson)

	Average/cod	CI95
3,5	38,68	2,34
AMR-WB1425	70,07	2,93
FCEKLT1425	63,46	2,96
FCMMDCT1425	39,25	3,53
Ref	100,00	0,00

Table 5: Global Results on all items (14.25kbps – Ericsson)

2. Medium bit rate: 24kbps (Ericsson)

Music	Av/codecs/ite	S	CI95
3,5	36,62		1,846
AMR-WB2385	74,48		2,874
FCEKLT24	62,33		2,865
FCMMDCT24	48,71		3,639
G729124	60,00		3,472
Ref	100,00		0

Table 6: Results on Music items (24kbps – Ericsson)

Speech	Av/codecs/ite	S	CI95
3,5	35,11		1,8
AMR-WB2385	83,39		2,659
FCEKLT24	83,04		2,584
FCMMDCT24	48,89		4,266
G729124	77,32		3,274
Ref	100,00		0

Table 7: Results on Speech items (24kbps – Ericsson)

Speech+Music	Av/codecs/ite	S	CI95
3,5	35,90		1,647
AMR-WB2385	89,43		1,735
FCEKLT24	79,95		2,287
FCMMDCT24	60,29		4,414
G729124	77,81		3,628
Ref	100,00		0

Table 8: Results on Speech + Music items (24kbps – Ericsson)

Noisy Speech	Av/codecs/ite	S	CI95
3,5	35,50		1,921
AMR-WB2385	93,14		2,056
FCEKLT24	92,07		2,301
FCMMDCT24	72,50		3,49
G729124	93,21		2,713
Ref	100,00		0

Table 9: Results on Noisy speech items (24kbps – Ericsson)

	Average/cod	S	CI95
3,5	35,75		1,767
AMR-WB2385	84,30		2,778
FCEKLT24	78,60		3,32
FCMMDCT24	55,63		4,406
G729124	75,76		4,003
Ref	100,00		0
Music	**Av/codecs/ite**	**S**	**CI95**

Table 10: Global Results on all items (24kbps – Ericsson)

3. High bit rate: 32kbps (Ericsson)

Music	Av/codecs/ite	S	CI95
3,5	35,28		2,17
AMR-WBP32	96,44		0,96
FCEKLT32	85,56		3,29
G722132	97,33		1,06
Ref	99,83		0,13

Table 11: Results on Music items (32kbps – Ericsson)

Speech	Av/codecs/ite	S	CI95
3,5	37,29		1,82
AMR-WBP32	95,17		1,7
FCEKLT32	94,50		1,37
G722132	77,42		3,35
Ref	99,92		0,08

Table 12: Results on Speech items (32kbps – Ericsson)

Speech+Mus	Av/codecs/ite	S	CI95
3,5	34,28		1,64
AMR-WBP32	97,50		1,29
FCEKLT32	96,06		1,3
G722132	96,78		1,22
Ref	99,89		0,09

Table 13: Results on Speech + Music items (32kbps – Ericsson)

Noisy Speecl	Av/codecs/ite	S	CI95
3,5	35,75		1,7
AMR-WBP32	99,67		0,17
FCEKLT32	98,25		0,55
G722132	97,42		0,87
Ref	99,67		0,22

Table 14: Results on Noisy speech items (32kbps – Ericsson)

	Average/cod	S	CI95
3,5	35,78		1,83
AMR-WBP32	96,82		1,28
FCEKLT32	93,28		2,11
G722132	91,26		2,74
Ref	99,85		0,13

Table 15: Global Results on all items (32kbps – Ericsson)

Annex 3: test results arrays - Nokia test site

4. Low bit rate: 14kbps (Nokia)

Music	Av/codec	CI95
3,5	30.94	8.68
AMR-WB1425	49.89	10.05
FCEKLT	38.67	8.79
FCRMDCT14	22.56	6.77
Ref	100	0

Table 1: Results on Music items (14.25kbps – Nokia)

Speech	Av/codec	CI95
3,5	34.38	7.4
AMR-WB1425	78.88	7.35
FCEKLT	74.63	7
FCRMDCT14	24.58	5.29
Ref	100	0

Table 2: Results on Speech items (14.25kbps – Nokia)

Speech+music	Av/codec	CI95
3,5	28.78	7.08
AMR-WB1425	61.61	8.51
FCEKLT	65.61	10.84
FCRMDCT14	37	9.4
Ref	100	0

Table 3: Results on Speech + Music items (14.25kbps – Nokia)

Noisy speech	Av/codec	CI95
3,5	33.33	11.24
AMR-WB1425	83	7.97
FCEKLT	69.92	5.73
FCRMDCT14	37.58	9.39
Ref	100	0

Table 4: Results on Noisy speech items (14.25kbps – Nokia)

	Av/codec	CI95
3,5	31.94	4.1172
AMR-WB1425	68	5.1972
FCEKLT	62.6	5.3635
FCRMDCT14	29.35	3.9702
Ref	100	0

Table 5: Global Results on all items (14.25kbps – Nokia)

5. Medium bit rate: 24kbps (Nokia)

Music	Av/codec	CI95
3,5	21.33	6.75
AMR-WB2385	78.06	6.66
FCEKLT24	74.56	9.21
FCRMDCT24	40	7.17
G729124	69.11	12.19
Ref	98.61	1.68

Table 6: Results on Music items (24kbps – Nokia)

Speech	Av/codec	CI95
3,5	28.08	6.49
AMR-WB2385	79	7.34
FCEKLT24	83.04	5.17
FCRMDCT24	46.17	10.06
G729124	79.46	5.88
Ref	99.46	0.601

Table 7: Results on Speech items (24kbps – Nokia)

Speech+music	Av/codec	CI95
3,5	23.33	6.81
AMR-WB2385	86.56	6.63
FCEKLT24	87.67	5.11
FCRMDCT24	61.39	8.04
G729124	85.11	6.99
Ref	97.78	2.343

Table 8: Results on Speech + Music items (24kbps – Nokia)

Noisy speech	Av/codec	CI95
3,5	27.75	8.18
AMR-WB2385	94	4.09
FCEKLT24	91.42	4.08
FCRMDCT24	62.75	13.74
G729124	86.25	8.42
Ref	96.83	3.593

Table 9: Results on Noisy speech items (24kbps – Nokia)

	Av/codec	CI95
3,5	25.15	3.5
AMR-WB2385	83.15	3.66
FCEKLT24	83.47	3.44
FCRMDCT24	51.19	5.24
G729124	79.42	4.44
Ref	98.39	0.96

Table 10: Global Results on all items (24kbps – Nokia)

6. High bit rate: 32kbps (Nokia)

Music	Av/codec	CI95
3,5	22.2	6.75
AMR-WBP32	95.67	6.66
FCEKLT32	82.87	9.21
FCRMDCT32	59.2	7.17
G722132	87.67	12.19
Ref	94.2	1.68

Table 11: Results on Music items (32kbps – Nokia)

Speech	Av/codec	CI95
3,5	25.7	6.49
AMR-WBP32	88.35	7.34
FCEKLT32	86.5	5.17
FCRMDCT32	63.75	10.06
G722132	78.05	5.88
Ref	97.95	0.601

Table 12: Results on Speech items (32kbps – Nokia)

Speech+music	Av/codec	CI95
3,5	22.47	6.81
AMR-WBP32	89.73	6.63
FCEKLT32	83.4	5.11
FCRMDCT32	71	8.04
G722132	90.87	6.99
Ref	94.8	2.343

Table 13: Results on Speech + Music items (32kbps – Nokia)

Noisy speech	Av/codec	CI95
3,5	26.5	8.18
AMR-WBP32	90.6	4.09
FCEKLT32	90.9	4.08
FCRMDCT32	79.4	13.74
G722132	86.3	8.42
Ref	99.3	3.593

Table 14: Results on Noisy speech items (32kbps – Nokia)

	Av/codec	CI95
3,5	24.15	4.22
AMR-WBP32	90.9	3.24
FCEKLT32	85.55	3.34
FCRMDCT32	67.03	5.72
G722132	85.03	3.56
Ref	96.45	2.34

Table 15: Global Results on all items (32kbps – Nokia)

Annex 4: Bit-rate results comparison for all codecs and types of items

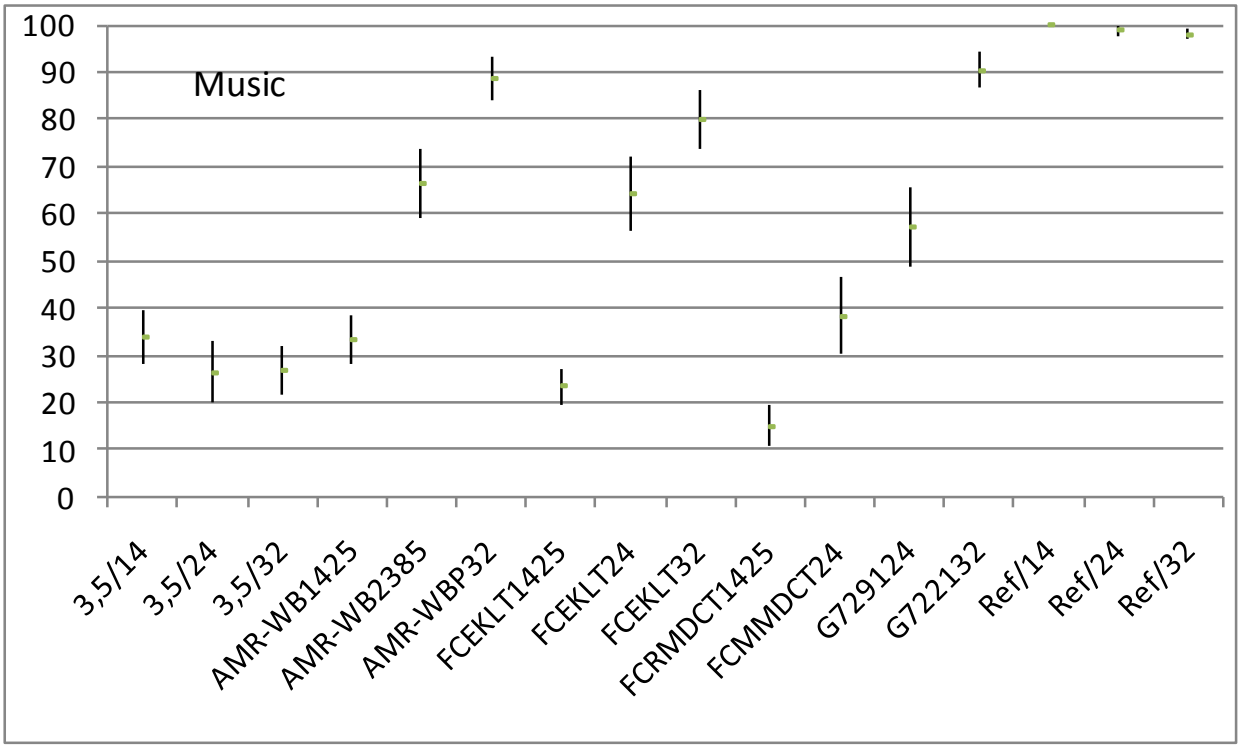


Figure 1: Results on Music items from 14kbps (left) to 32kbps (right) via 24kbps (middle) – Orange Labs

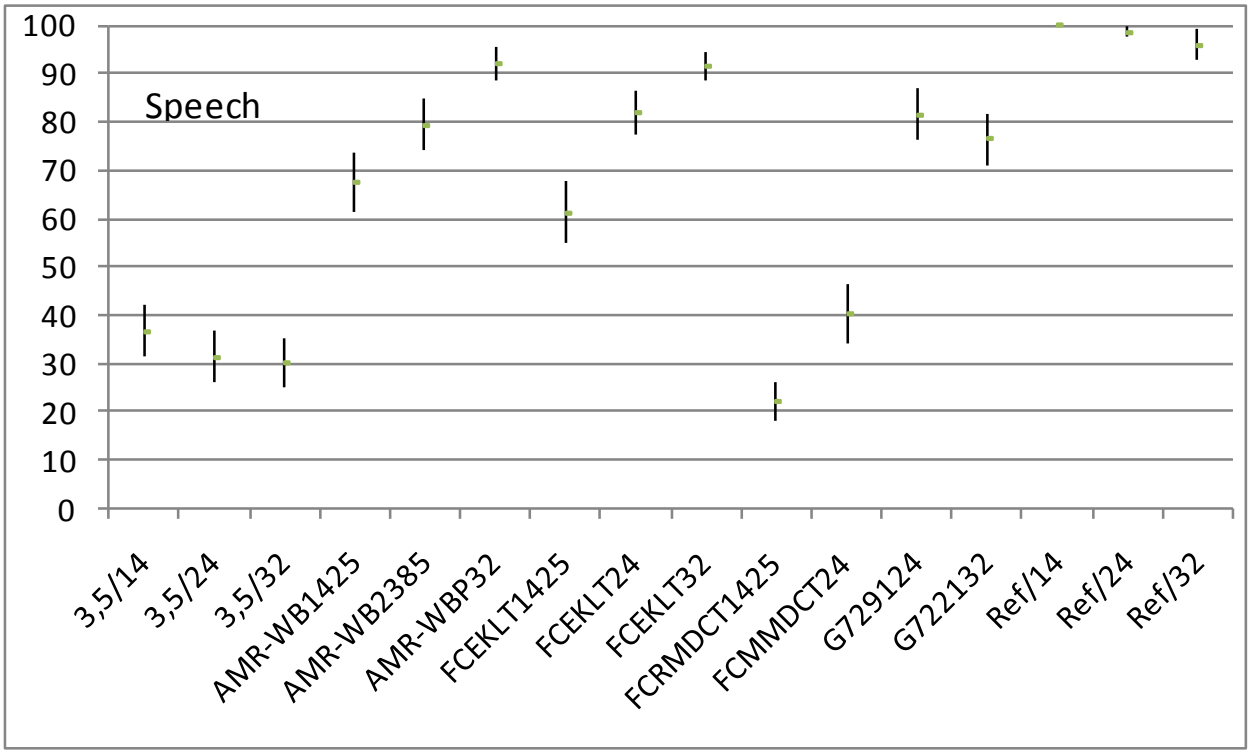


Figure 2: Results on Speech items from 14kbps (left) to 32kbps (right) via 24kbps (middle) – Orange Labs

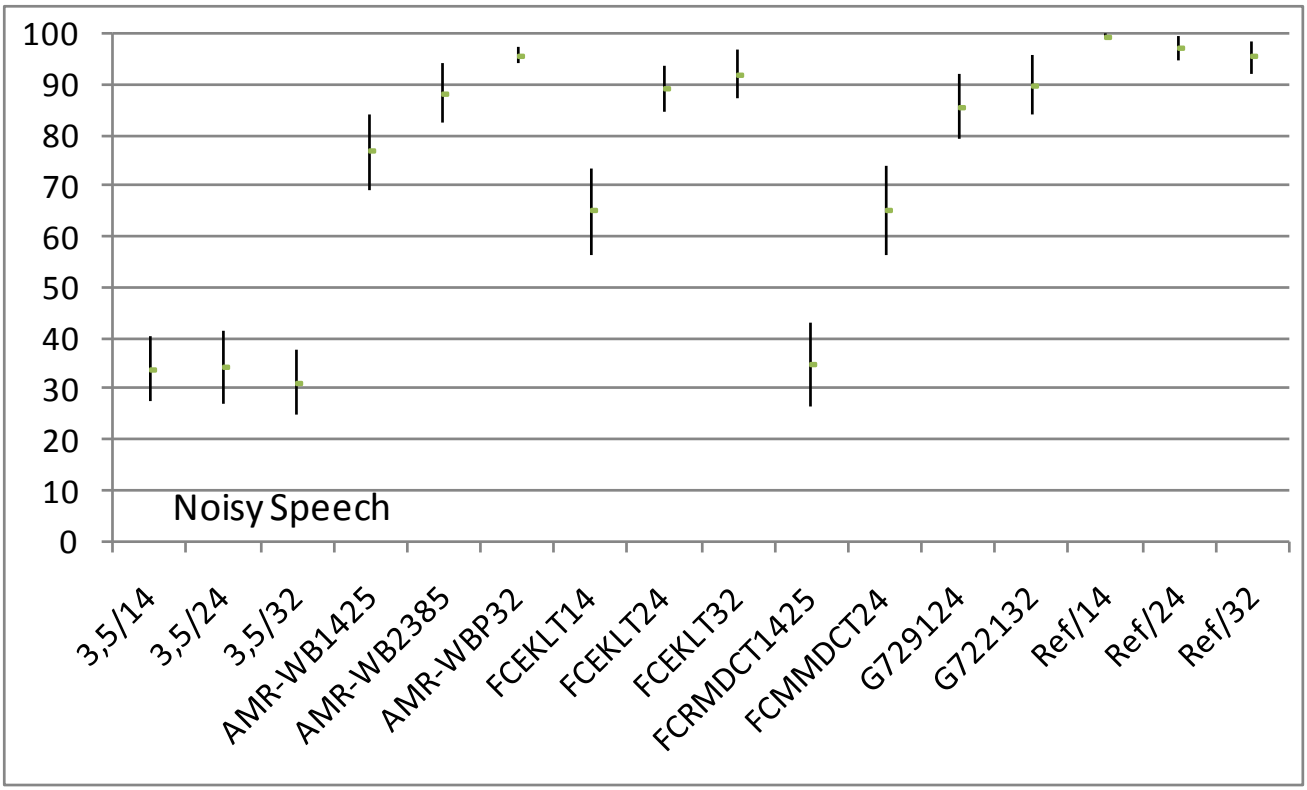


Figure 3: Results on Noisy Speech items from 14kbps (left) to 32kbps (right) via 24kbps (middle) – Orange Labs

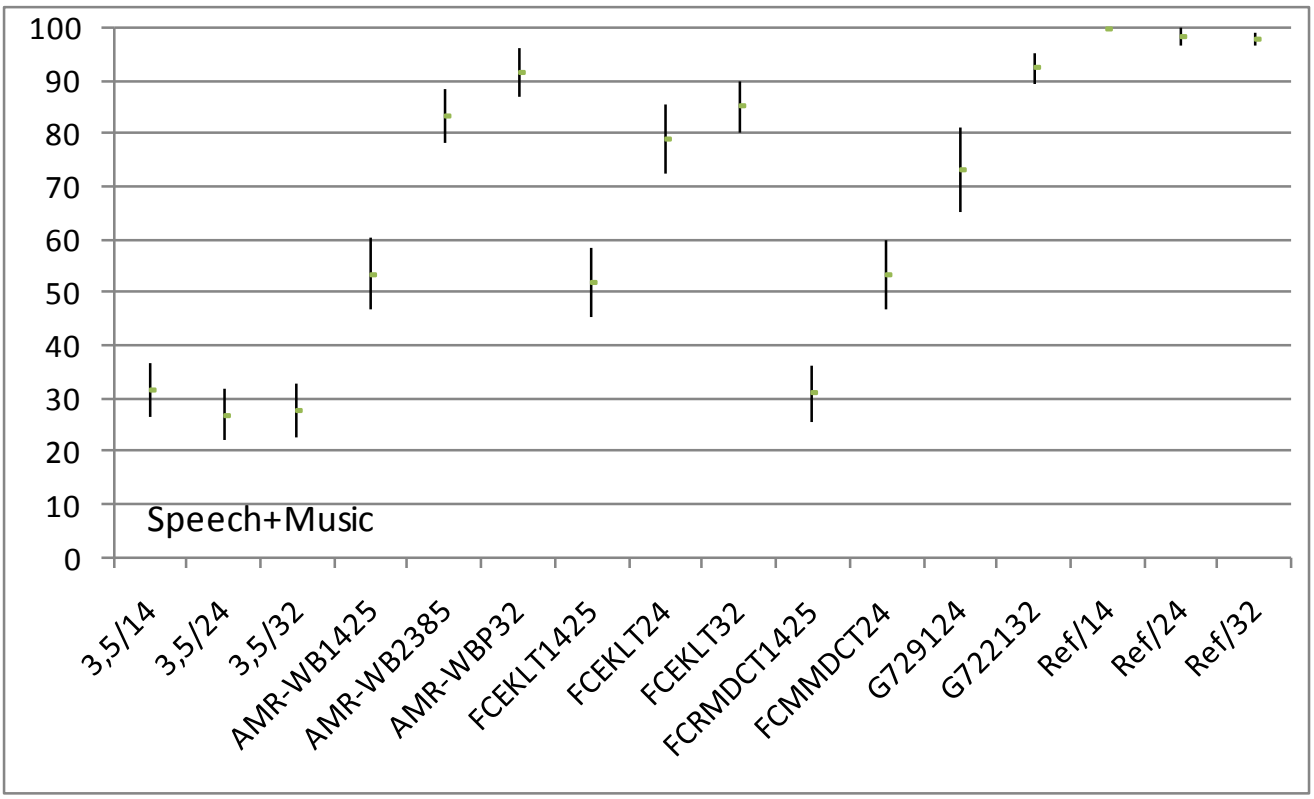


Figure 10: Results on Speech + Music items from 14kbps (left) to 32kbps (right) via 24kbps (middle) – Orange Labs