TSG-SA4#30 meeting *Tdoc S4 (04)0020*
February 23-27, 2004, Malaga, Spain

Source: TSG SA WG4 (France Telecom)
Title: Report on 3G conversation tests phase 2

Comparison of quality offered by different speech coders.

Document For: Approval
Agenda Item: 7.4.3

1 Introduction

3GPP SA4 has approved the test plans defined for comparing AMR Narrow and Wide-band Packet Switched codecs with other codecs, for different packet loss ratio (Tdoc S4-030747). Based on this test plan, France Telecom R&D has implemented the test bed. The general definition of the test bed is available in the Documents defined above. This new Document does not reproduce the contents of the defined above, but only presents the test results. The test was done in two different languages (French and Arabic). The table 1 below just reminds the 16 tested conditions.

Condition	**Experimental factors**		**Symbol in figures**
	IP conditions (Packet loss ratio)	***Mode***	
1	0%	AMR NB 6.7kbit/s	NB 6.7
2	0%	AMR-NB 12.2 kbit/s	NB 12.2
3	0%	AMR-WB 12.65 kbit/s	WB 12.65
4	0%	AMR-WB 15.85 kbit/s	WB 15.85
5	0%	G. 723.1 6.4 kbit/s	G 723.1
6	0%	G.729 8 kbit/s	G.729
7	0%	G.722 64 kbit/s + plc	G.722
8	0%	G.711 + plc	G.711
9	3%	AMR NB 6.7kbit/s	NB 6.7
10	3%	AMR-NB 12.2 kbit/s	NB 12.2
11	3%	AMR-WB 12.65 kbit/s	WB 12.65
12	3%	AMR-WB 15.85 kbit/s	WB 15.85
13	3%	G. 723.1 6.4 kbit/s	G 723.1
14	3%	G.729 8 kbit/s	G.729
15	3%	G.722 64 kbit/s + plc	G.722
16	3%	G.711 + plc	G.711

Table 1: Tested conditions (plc = packet loss concealment)

2 Test results

2.1 French language

The figure 1 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the Global Quality criterion for the sixteen conditions. It appears that the Mean Opinion Scores obtained for 3 % of packet losses are systematically less important than those obtained for 0 % of packet losses, except for the codec G722 and G711. One also observes some judgment differences between the different codecs for a same packet loss ratio, the more appreciated being the wide-band codecs.
A Variance Analysis ANOVA shows that there is a weak effect of the packet loss ration $F(1,31) = 4.93$ $p<0.05$), and a weak but very significant effect of the codec ($F(7,217) = 4.05$ $p<0.0001$), as well as an interaction between the Packet loss ratio and the codec ($F(7,217) = 2.18$ $p<0.05$).

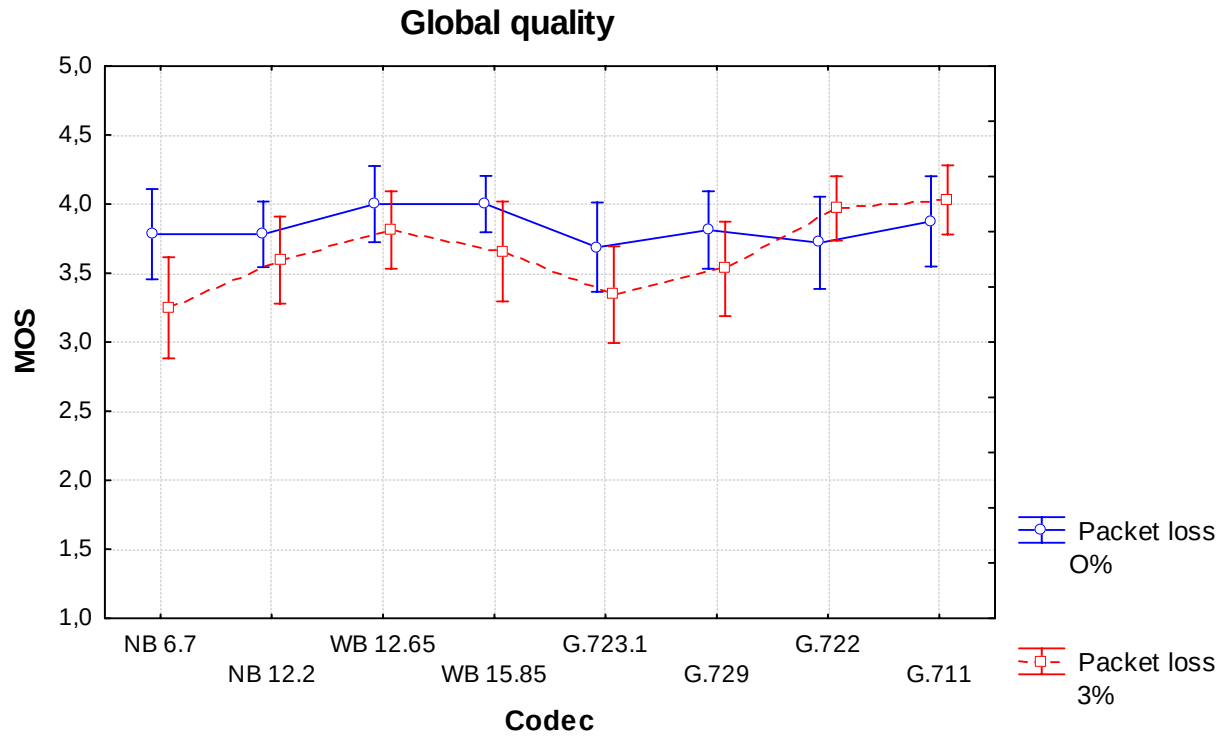

Figure 1: Mean Opinion Scores obtained with the Global Quality Criterion according to the Codec conditions, and the IP Packet loss ratio.

Table 2 below shows the significance of the differences between the different codecs, the two packet loss ratio considered, according to a Tukey test. The significant differences are marked with a star.

	AMR NB 6.7	AMR NB 12.2	AMR WB 12.65	AMR WB 15.85	G.723.1 6.4	G.729 8	G.722 64+plc	G.711 plc
AMR NB 6.7		0,83	0,02 *	0,14	1,00	0,89	0,10	0,01 *
AMR NB 12.2	0,83		0,59	0,94	0,83	1,00	0,89	0,33
AMR WB 12.65	0,02 *	0,59		1,00	0,02 *	0,50	1,00	1,00
AMR WB 15.85	0,14	0,94	1,00		0,14	0,89	1,00	0,97
G.723.1 6.4	1,00	0,83	0,02 *	0,14		0,89	0,10	0,01 *
G.729 8	0,89	1,00	0,50	0,89	0,89		0,83	0,26
G.722 64+plc	0,10	0,89	1,00	1,00	0,10	0,83		0,98
G.711 plc	0,01 *	0,33	1,00	0,97	0,01 *	0,26	0,98	

Table 2: Significance of the differences between codecs

The table 3 below gives the correlation coefficients between the different criteria. They are all significant although not very correlated. Therefore, the effects obtained with the four other criteria are rather similar to those obtained with the Global quality criterion. There are succinctly given in the following.

	c1	c2	c3	c4	c5
c1: Voice quality	1.00	0.47	0.44	0.47	0.58
c2: Understanding	0.47	1.00	0.55	0.51	0.56
c3: Interaction	0.44	0.55	1.00	0.50	0.60
c4: Defaults perception	0.47	0.51	0.50	1.00	0.71
c5: Global quality	0.58	0.56	0.60	0.71	1.00

Table 3: Correlation coefficients between the five criteria

For the **Voice quality criterion**, the effects are similar than those obtained with the Global quality criterion: effect of the packet loss except for the G. 722 and the G. 711, effect of the codec. Figure 2 illustrates these effects with the Mean Opinion Scores and the associated confidence intervals (95%) obtained with for the sixteen conditions. Annex 1 gives the results of the different ANOVA conducted on scores for each criterion.

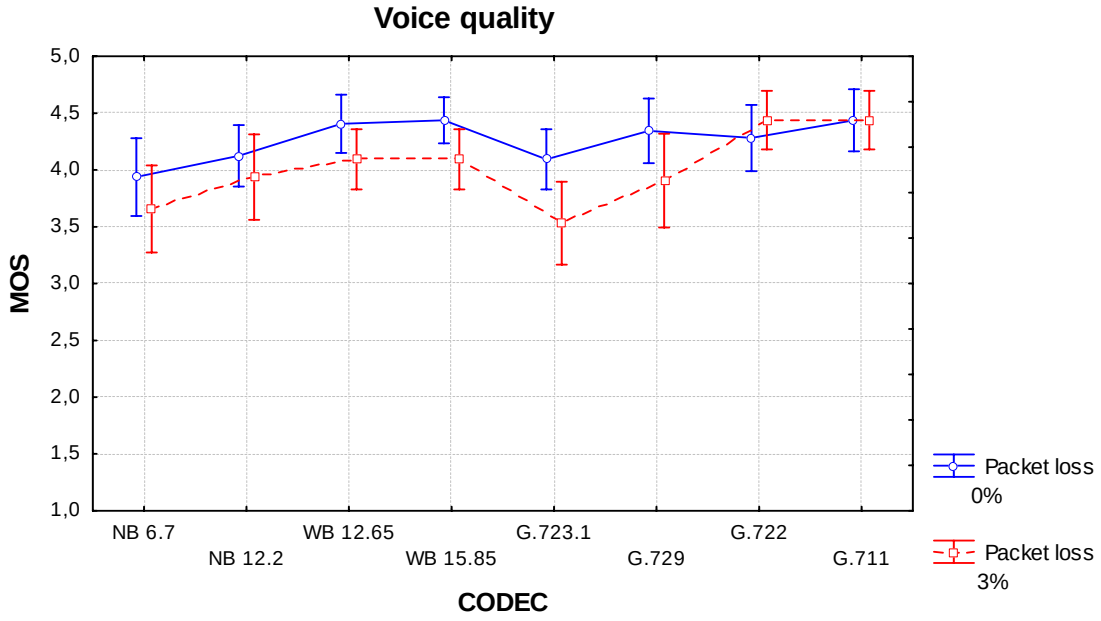

Figure 2: Mean Opinion Scores obtained with the Voice quality criterion according to the Codec conditions, and the IP Packet loss ratio.

The figure 3 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the **Understanding criterion** for the sixteen conditions. The results of the ANOVA are given in annex 1. Notice that the MOS are really good (superior to 4 MOS) and there is no effect of the packet loss ratio ($F(1,31) = 3.96$ $p=0.055$). Therefore, even with some packet losses, the eight tested codecs offer a good intelligibility.

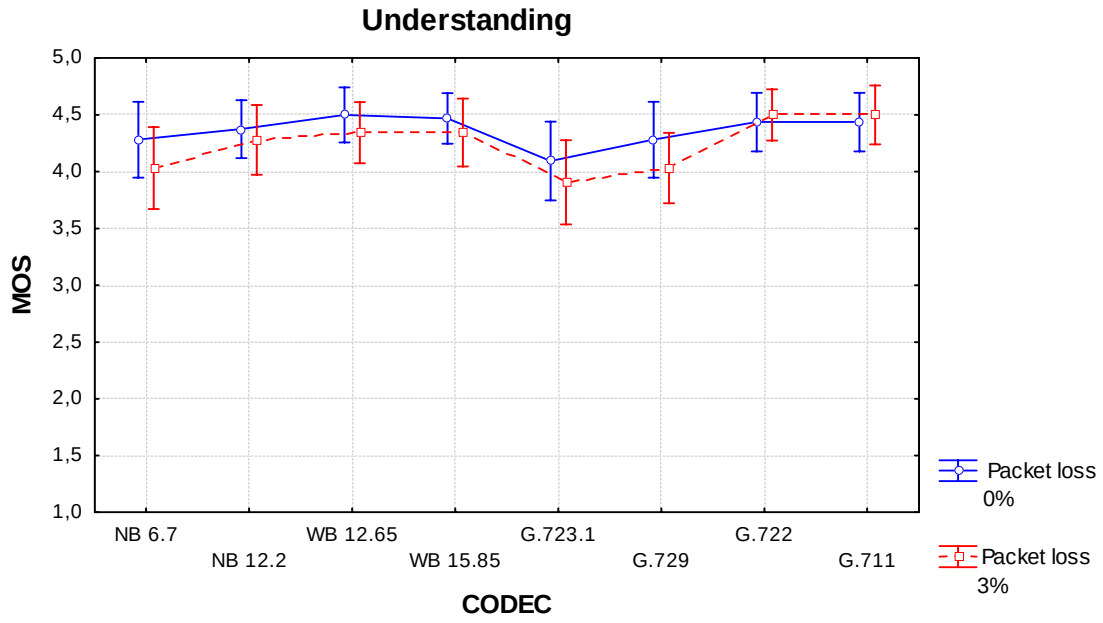

Figure 3: Mean Opinion Scores obtained with the Understanding criterion according to the Codec conditions, and the IP Packet loss ratio.

The figure 4 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the **Interaction criterion** for the sixteen conditions. It appears that the MOS differences between the codecs are weak. The results of the ANOVA are given in annex 1 (the effect of the codec becomes very weak for this criterion: $F(7,217) = 2.39$ $p<0.05$). Since the intelligibility is ensured (cf. understanding criterion), the delay becomes one of the most relevant parameter susceptible to influence the interaction. The delay was the same for all the codecs (about 300 ms). It appears that this delay has the same effect for all the tested codecs.

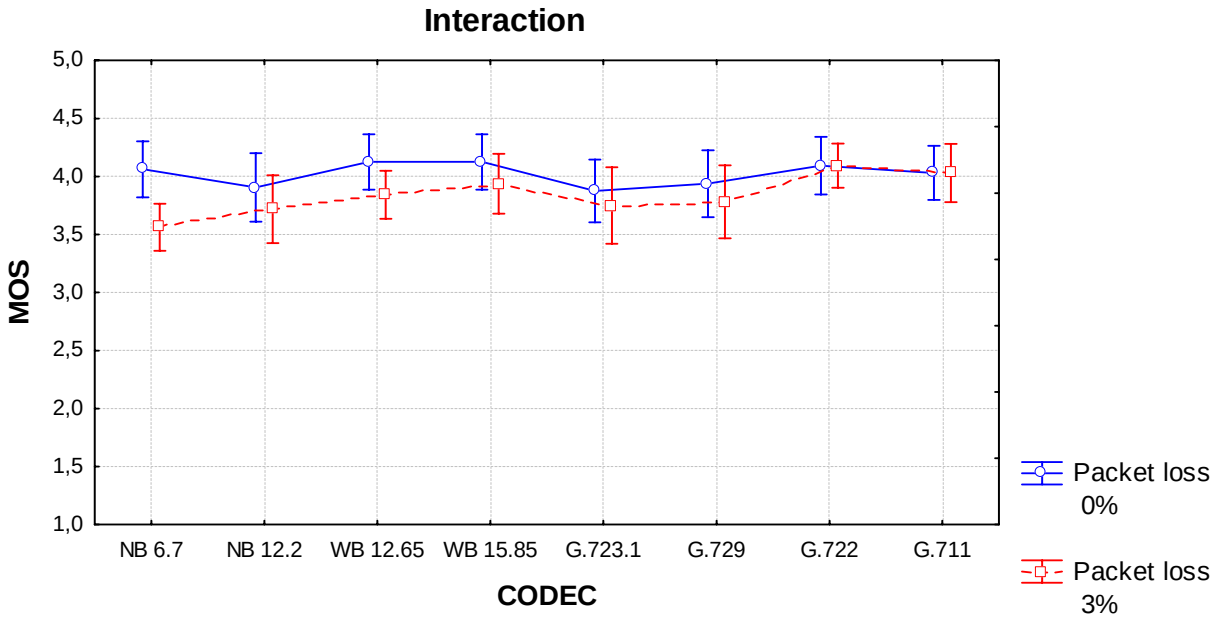

Figure 4: Mean Opinion Scores obtained with the Interaction criterion according to the Codec conditions, and the IP Packet loss ratio.

Finally, the figure 5 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the **Defaults perception criterion** for the sixteen conditions.

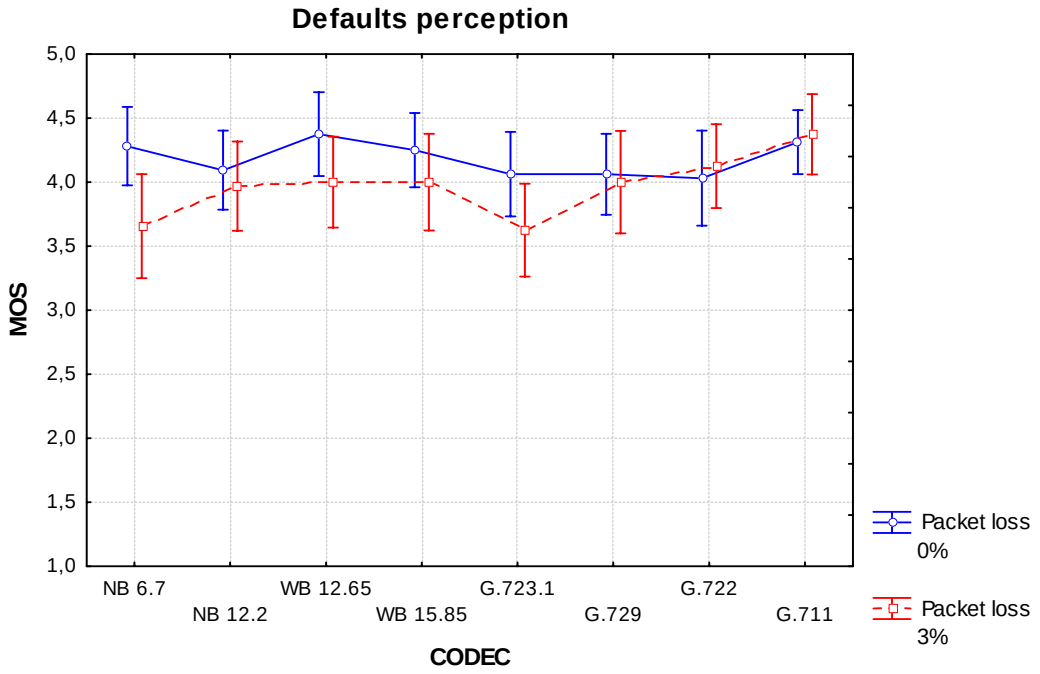

Figure 5: Mean Opinion Scores obtained with the Defaults perception criterion according to the Codec conditions, and the IP Packet loss ratio.

2.2 Arabic language

The figure 6 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the Global Quality criterion for the sixteen conditions. It appears that the Mean Opinion Scores obtained for 3 % of packet losses are systematically less important than those obtained for 0 % of packet losses, except for the codec G722 and G711. One also observes some judgment differences between the different codecs for a same packet loss ratio.

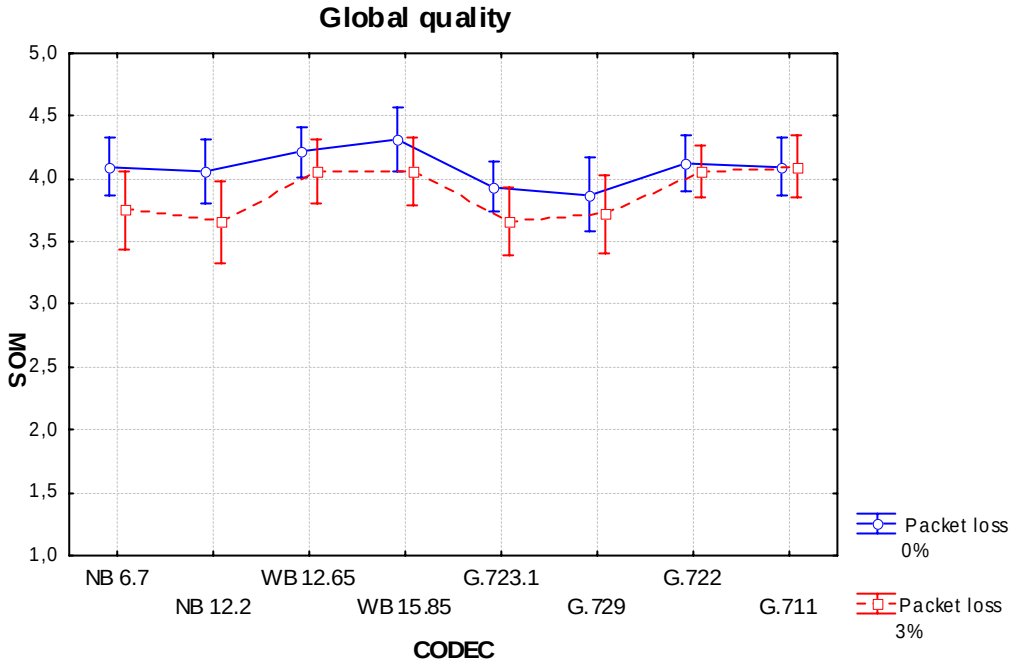

Figure 6: Mean Opinion Scores obtained with the Global Quality Criterion according to the Codec conditions, and the IP Packet loss ratio.

A Variance Analysis ANOVA confirms that there is an effect of the packet loss ration $F(1,31) = 18$ $p<0.0001$, as well as an effect of the codec ($F(7,217) = 4.09$ $p<0.0001$), but no interaction between the Packet loss ratio and the codec ($F(7,217) = 0.81$ $p=0.58$).

Table 4 below shows the significance of the differences between the different codecs, the two packet loss ratio considered, according to a Tukey test. The significant differences are marked with a star.

	AMR NB 6.7	***AMR NB 12.2***	***AMR WB 12.65***	***AMR WB 15.85***	***G.723.1 6.4***	***G.729 8***	***G.722 64+plc***	***G.711 plc***
AMR NB 6.7		1.00	0.51	0.26	0.95	0.95	0.79	0.79
AMR NB 12.2	1.00		0.19	0.07	1.00	1.00	0.42	0.42
AMR WB 12.65	0.51	0.19		1.00	0.04*	0.04*	1.00	1.00
AMR WB 15.85	0.26	0.07	1.00		0.01*	0.01*	0.99	0.99
G.723.1 6.4	0.95	1.00	0.04	0.01		1.00	0.14	0.14
G.729 8	0.95	1.00	0.04	0.01	1.00		0.14	0.14
G.722 64+plc	0.79	0.42	1.00	0.99	0.14	0.14		1.00
G.711 plc	0.79	0.42	1.00	0.99	0.14	0.14	1.00	

Table 4: Significance of the differences between codecs

The table 5 below gives the correlation coefficients between the different criteria. They are all significant although not very correlated. Therefore, the effects obtained with the four other criteria are rather similar to those obtained with the Global quality criterion. There are succinctly given in the following.

	c1	c2	c3	c4	c5
c1: Voice quality	1.00	0.46	0.56	0.49	0.58
c2: Understanding	0.46	1.00	0.51	0.32	0.46
c3: Interaction	0.56	0.51	1.00	0.52	0.62
c4: Defaults perception	0.49	0.32	0.52	1.00	0.50
c5: Global quality	0.58	0.46	0.62	0.50	1.00

Table 5: Correlation coefficients between the five criteria

The figure 7 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the **Voice quality criterion** for the sixteen conditions. Annex 1 gives the results of the different ANOVA conducted on scores for each criterion. It appears that there are effects of the packet loss ratio and of the codec. According to the significant interaction between the two factors, the effect of the packet loss ratio depends on the codec: the WB 12.65, the G.729, the G.722, and the G. 711 seem to have a good behavior in presence of packet losses.

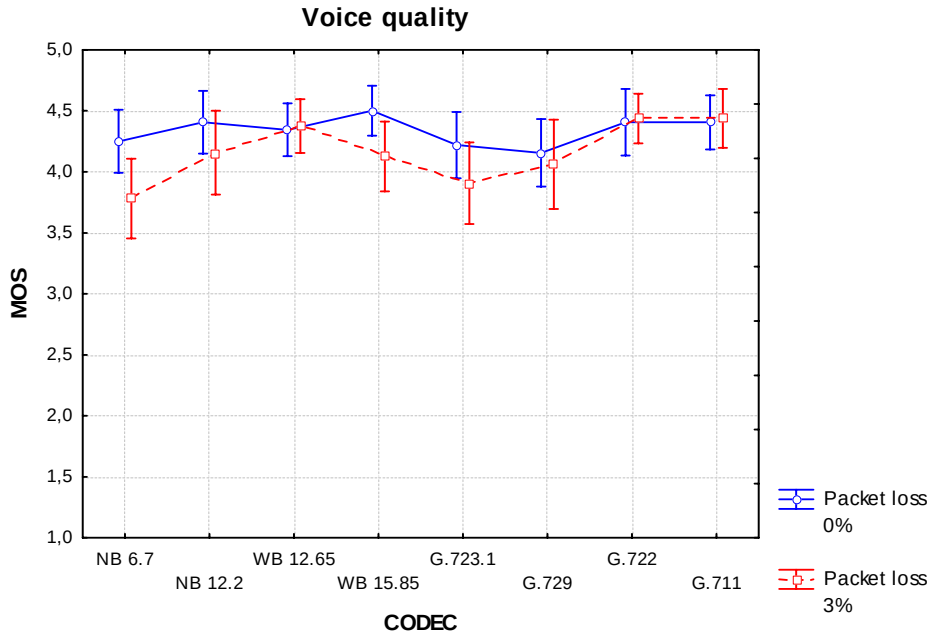

Figure 7: Mean Opinion Scores obtained with the Voice quality criterion according to the Codec conditions, and the IP Packet loss ratio.

The figure 8 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the **Understanding criterion** for the sixteen conditions. It appears that, even in presence of packet losses, the scores are very good, so a good intelligibility is ensured for all the codecs.

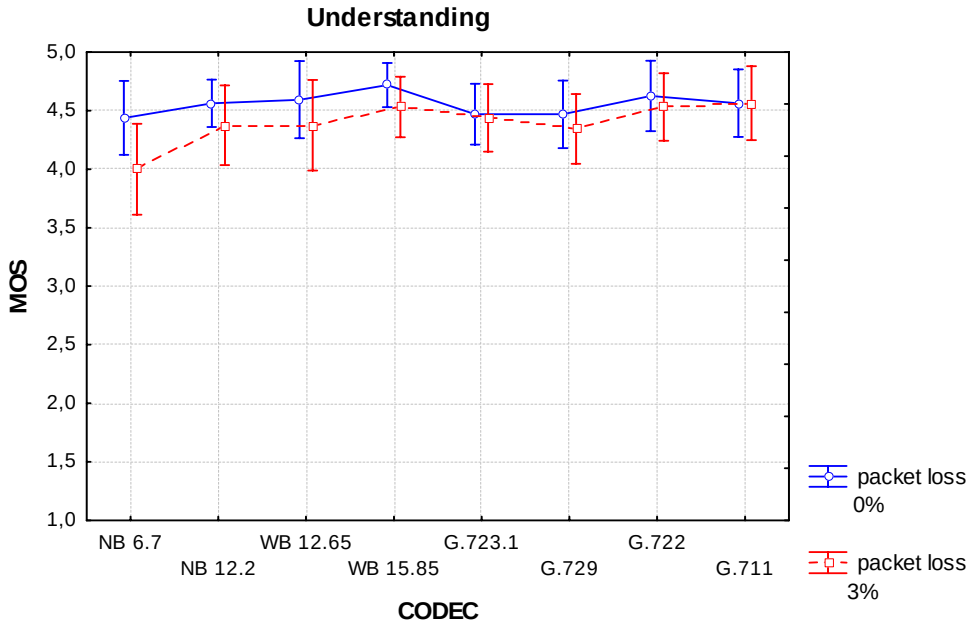

Figure 8: Mean Opinion Scores obtained with the Understanding criterion according to the Codec conditions, and the IP Packet loss ratio.

The figure 9 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the **Interaction criterion** for the sixteen conditions.

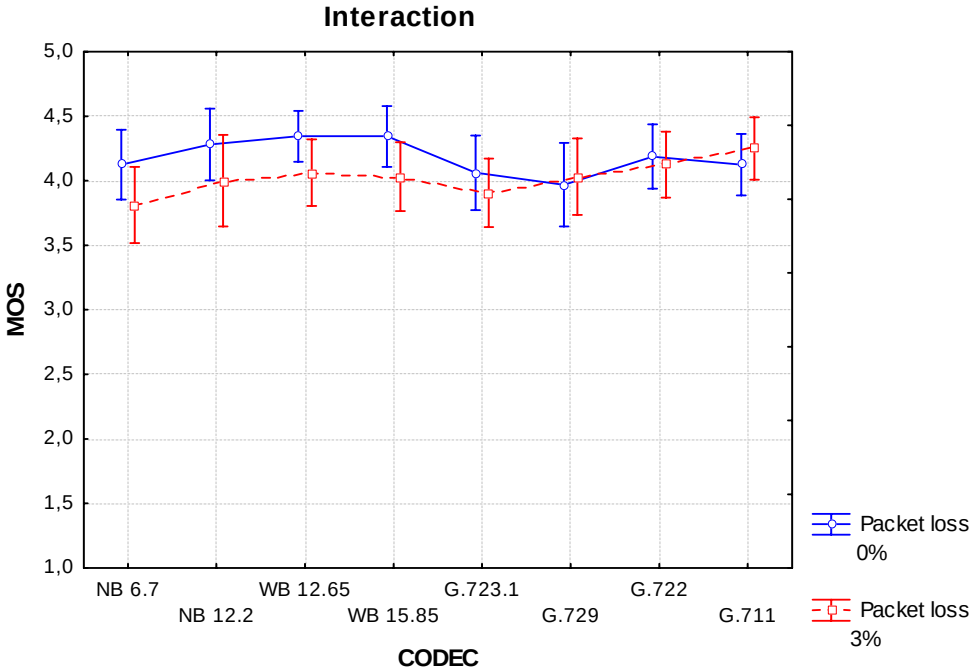

Figure 9: Mean Opinion Scores obtained with the Interaction criterion according to the Codec conditions, and the IP Packet loss ratio.

Finally, the figure 10 below shows the Mean Opinion Scores and the associated confidence intervals (95%) obtained with the **Defaults perception criterion** for the sixteen conditions.

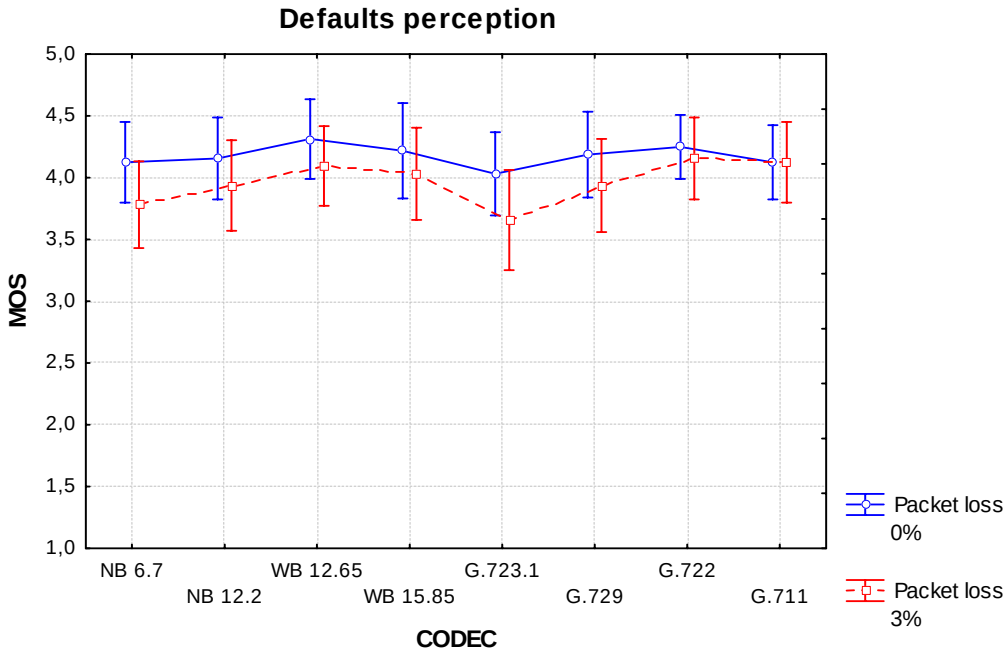

Figure 10: Mean Opinion Scores obtained with the Defaults perception criterion according to the Codec conditions, and the IP Packet loss ratio.

2.3 Comparison French / Arabic

The figure 11 below shows the MOS and the associated confident intervals obtained for the criterion Global quality, according to the codec, the packet loss, and the language. It appears that, especially in absence of packet losses, the MOS obtained with the Arabic language are superior to those obtained with the French language. An inter-group variance analysis confirms that there is a weak effect of the language ($F(1,62) = 4.92$ $p<0.05$). This effect is less marked with 3% of packet losses.

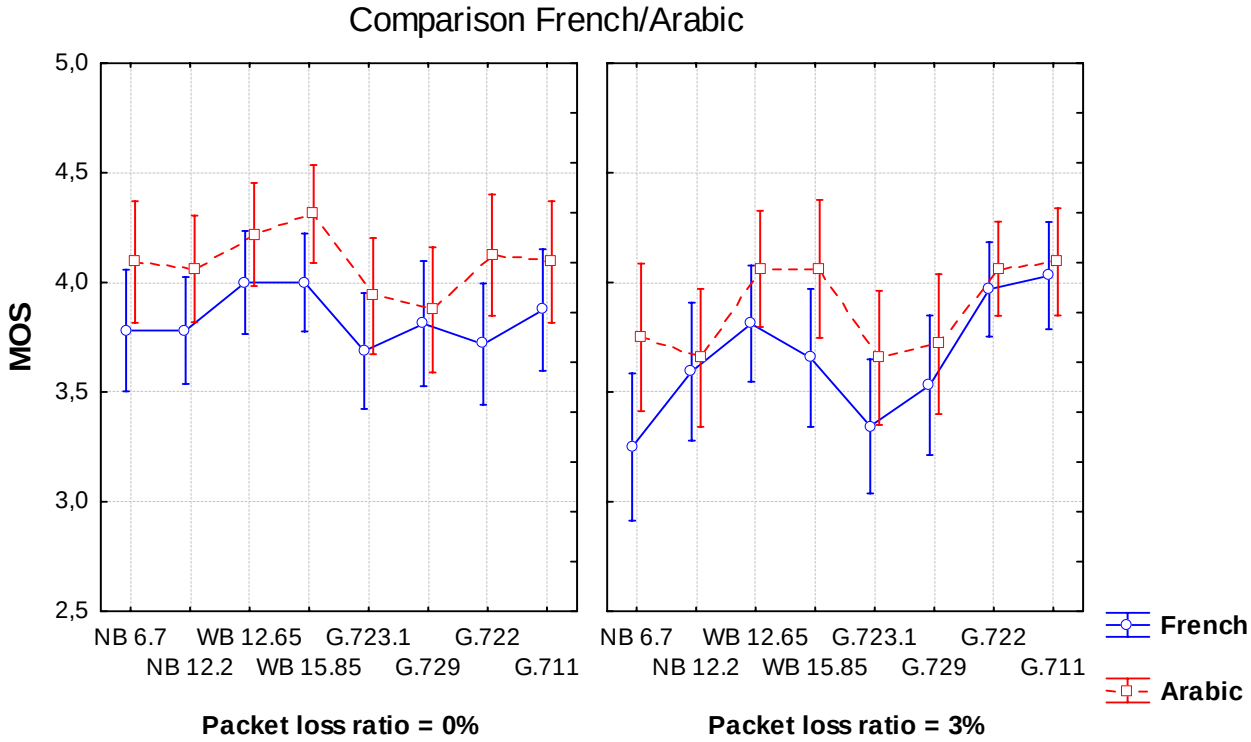

Figure 11: Comparison between the two languages.

3 Conclusion

Apart for G.711 and G.722, subjects are sensible to packet loss ratio and it appears that the Mean Opinion Scores obtained for 3 % of packet losses are systematically less important than those obtained for 0 % of packet losses. It can be noted that good intelligibility is insured for the eight codecs under test.

It appears that in most of the cases the differences between codecs are not significant, nevertheless some judgment differences between the different codecs for a same packet loss ratio can be observed, the more appreciated codecs being wide band codecs. The difference is observable but not very significant in quiet environment.

ANNEX 1

Results of the difference Variance Analysis conducted on scores for the different criteria, for French language

The considered factors are the packet loss ratio (two levels: 0%, 3%) and the codec (eight levels)

Criterion	Packet loss ratio effect	Codec effect	Interaction packet loss ratio/codec
Global quality	$F(1,31) = 4.93$ $p<0.05$	$F(7,217) = 4.05$ $p<0.0001$	$F(7,217) = 2.18$ $p<0.05$
Voice quality	$F(1,31) = 10.25$ $p<0.005$	$F(7,217) = 7.42$ $p<0.0001$	$F(7,217) = 0.86$ $p=0.122$
Understanding	$F(1,31) = 3.96$ $p=0.055$	$F(7,217) = 4.8$ $p<0.0001$	$F(7,217) = 0.63$ $p=0.728$
Interaction	$F(1,31) = 7.94$ $p<0.05$	$F(7,217) = 2.39$ $p<0.05$	$F(7,217) = 1.12$ $p=0.34$
Defaults perception	$F(1,31) = 5.9$ $p<0.05$	$F(7,217) = 1.4$ $p<0.05$	$F(7,217) = 1.54$ $p=0.15$

Results of the difference Variance Analysis conducted on scores for the different criteria, for Arabic language

The considered factors are the packet loss ratio (two levels: 0%, 3%) and the codec (eight levels)

Criterion	Packet loss ratio effect	Codec effect	Interaction packet loss ratio/codec
Global quality	$F(1,31) = 18$ $p<0.0001$	$F(7,217) = 4.09$ $p<0.0001$	$F(7,217) = 0.81$ $p=0.58$
Voice quality	$F(1,31) = 9.53$ $p<0.005$	$F(7,217) = 4.75$ $p<0.0001$	$F(7,217) = 2.171$ $p<0.05$
Understanding	$F(1,31) = 21.87$ $p<0.0001$	$F(7,217) = 2.72$ $p<0.05$	$F(7,217) = 0.29$ $p= 1.23$
Interaction	$F(1,31) = 8.69$ $p<0.05$	$F(7,217) = 2.13$ $p<0.05$	$F(7,217) = 1.41$ $p=0.199$
Defaults perception	$F(1,31) = 7.9$ $p<0.05$	$F(7,217) = 1.98$ $p=0.059$	$F(7,217) = 0.52$ $p=0.81$