Joint TSG-S4#10 - SMG11#15 Meeting
Helsinki, Finland **S4/SMG11 Tdoc 137/00**
February 28th - March 3rd 2000

Title: **Draft for the PC experiment and proposed DCR testing for the NS test plan**

Version: **0.1**

Source: **France Telecom /CNET**[1]

This document describes the proposed procedures for conducting two experiments of the subjective test plan for assessing the minimum performance specifications of any AMR noise suppression solution in the future. These experiment are designed to verify the requirements : No degradation in clean speech (Exp#1) and No artifacts in residual noise, and No speech clipping and no reduction in intelligibility (Exp#2), according to the the associated Section in Stage 1 Description (TS GSM 02.76 v. 2.0.0). Exp#1 is derived from the similar PC test included in the Test plan Specification for the AMR-NS Selection Phase (Exp#2 in Tdoc SMG11/S4 356/99 R3, [1]). Exp#2 is made up of three sub-experiments, one for each of the three types of noise tested (car, street and babble). The rationale to design a Modified DCR test for the second experiment is clarified in section II.1.

In the following an overview of two experiments of the final test plan is given.

Experiment #1	**Degradation in Clean Speech (Pair Comparison Test)** No degradation in clean speech.
Experiment #2 **sub-exp. #2a** (car noise) **sub-exp. #2b** (street noise) **sub-exp. #2c** (babble noise)	**Artefacts and Clipping Effects in Background Noise Conditions (Modified DCR)** No artefacts in residual noise, and No speech clipping and no reduction in intelligibility.

The Absolute Category Rating (ACR) and the Degradation Category Rating (DCR) methods are described in ITU-T Recommendation P.800 [2].

1

Dominique Pascal
France Télécom / CNET
DIH/EQS
2 Avenue Pierre Marzin
Technopole Anticipa
22307 Lannion Cedex
FRANCE
Tel: +33 2 96 05 15 78
Fax: +33 2 96 05 13 16
E-mail: dominique.pascal@cnet.francetelecom.fr

I. Experiment 1 : Degradation in Clean Speech (Pair Comparison Test)

I.1 Introduction

This PC (Paired-Comparison) experiment was prepared to test requirements 4.5.1.2 in the associated Section in Stage 1 Description (TS GSM 02.76 v. 2.0.0), i.e. **No degradation in clean speech**. This PC experiment will be run for the whole set of bit rates of the base vocoder, in single and tandem connection.

The test methodology is direct, paired, forced choice comparison (i.e. A versus B test method with forced choice) . The question that we are trying to answer with this test is not "What is the rank order of several coders?" but rather "Does the quality of coder with noise suppression (+NS) meet or exceed the quality of the coder without NS for a given condition?". The direct comparison A/B test methodology can answer this question by considering the proportion (or percent) of the measures where the candidate was preferred over the standard. Each individual judgement is a binary decision. A rank order approach could be taken as noted in the Handbook of Telephonometry [3] regarding Paired Comparisons but notes: "In the scaling modulus is included the common standard deviation, which is, however, unknown and so does not permit calculating confidence limits for the scale positions obtained."

Unlike a simple mean where the Standard Error (SE) does not depend on the true mean value, the SE of a measured proportion P does depend on the true proportion p. From a sample we have an estimate of p (P) and so the true SE is unknown unless p is known. However, when samples are large and you can assume the sampling distribution of proportions are approximately normal, the 100 (1-alpha) % confidence limits are given by the following:

$$(N/(N+z^2))(P+z^2/2N) - z\sqrt{PQ/N + z^2/(4N^2)}$$

where N is the number of measures, z is the standard score in a normal distribution cutting off the lower alpha/2 proportion of cases, P is the proportion of the measures where a coder is preferred, and Q is 1-P. For the A/B experiment proposed here, with 24 subjects each making two independent measures (A/B and B/A) of the preference of the candidate coder over the standard coder for four talkers (two male and two female) each condition and with one repeat , the effective N is 384. In order to accommodate the repeat measure, single sentence samples will be used. This provides the additional benefit of directly adjacent A/B comparisons during presentation. The repeat measure will be made using a unique second sentence.

I.2. Test Factors and Conditions

The PC test will be run for the following basic vocoder conditions:

- Bit Rates of 4.75 kbit/s, 5.15 kbit/s, 5.9 kbit/s, 6.7 kbit/s, 7.4 kbit/s, 7.95 kbit/s, 10.2 kbit/s and 12.2 bit/s.
- Single codec and Codec/Codec Tandem.

This results in a single PC experiment with clean source speech and no channel impairments. The speech material used in these experiments are 4s samples (single sentence).

The following table shows the testing factors to be used in this experiment. A list of test conditions is given in Table 1.2.

Main Codec Conditions	**#**	**Notes**
Noise Suppresser Candidate	1	
Codec	1	AMR
Codec Modes (FR/HR)	HR	All 8 AMR modes
	FR	
BERs	0	Clear channel, no transmission errors
Input level	1	nominal: -26dB relative to OVL
Acoustic Background Noise	0	None
Tandeming	1	Self tandem condition
Input Characteristic	1	GSM Filtered
Codec references	***#***	***Notes***
Test vocoders	1	AMR with NS
Reference vocoder	6	AMR at 12.2, 10.2, 7.4, 6.7, 5.9 & 5.15
Other references	#	Notes
Direct		nominal level, GSM Filtered
MNRU	0	None, but used in preliminaries
Ideal Noise Suppression	0	None
Common Conditions	***#***	***Notes***
GSM Channel	0	NO channel model
Number of talkers	4	2 male + 2 female
Number of speech samples	52	12/talker + 1 practice/talker
Sentences/sample	1	Single sentence stimuli
Listening Level	1	-15dBPa (79dB SPL) at ERP
Listeners	24	Naive Listeners
Randomizations	6	6 groups of 4 listeners
Rating Scale	1	PC Instructions
Replications	2	Original Presentation + repeat w/ 2nd sentence

Table 1.1: **Factors and conditions for Experiment 1**

I.3. Test Conditions for Experiment 1

Cond.	Reference Codec	Processed Codec	Trans-codings	Speech sample number (6 sequences)
1	AMR@12.2	AMR@12.2	1	2 3 4 5 6 1
2	AMR@12.2	AMR@12.2	2	3 4 5 6 1 2
3	AMR@10.2	AMR@10.2	1	1 2 3 4 5 6
4	AMR@10.2	AMR@10.2	2	4 5 6 1 2 3
5	AMR@7.4	AMR@7.4	1	5 6 1 2 3 4
6	AMR@7.4	AMR@7.4	2	6 1 2 3 4 5
7	AMR@6.7	AMR@6.7	1	2 3 4 5 6 1
8	AMR@6.7	AMR@6.7	2	3 4 5 6 1 2
9	AMR@5.9	AMR@5.9	1	1 2 3 4 5 6
10	AMR@5.9	AMR@5.9	2	4 5 6 1 2 3
11	AMR@5.15	AMR@5.15	1	5 6 1 2 3 4
12	AMR@5.15	AMR@5.15	2	6 1 2 3 4 5
13	AMR@12.2	AMR/NS@12.2	1	2 3 4 5 6 1
14	AMR@12.2	AMR/NS@12.2	2	3 4 5 6 1 2
15	AMR@10.2	AMR/NS@10.2	1	1 2 3 4 5 6
16	AMR@10.2	AMR/NS@10.2	2	4 5 6 1 2 3
17	AMR@7.95	AMR/NS@7.95	1	5 6 1 2 3 4
18	AMR@7.95	AMR/NS@7.95	2	6 1 2 3 4 5
19	AMR@7.4	AMR/NS@7.4	1	5 6 1 2 3 4
20	AMR@7.4	AMR/NS@7.4	2	6 1 2 3 4 5
21	AMR@6.7	AMR/NS@6.7	1	2 3 4 5 6 1
22	AMR@6.7	AMR/NS@6.7	2	3 4 5 6 1 2
23	AMR@5.9	AMR/NS@5.9	1	1 2 3 4 5 6
24	AMR@5.9	AMR/NS@5.9	2	4 5 6 1 2 3
25	AMR@5.15	AMR/NS@5.15	1	5 6 1 2 3 4
26	AMR@5.15	AMR/NS@5.15	2	6 1 2 3 4 5
27	AMR@4.75	AMR/NS@4.75	1	5 6 1 2 3 4
28	AMR@4.75	AMR/NS@4.75	2	6 1 2 3 4 5
29-56	Reversed order of the reference and processed speech samples in cond. 1-28			
57 - 84	Repeat of conditions 1 – 28 with Speech Sample Number +6			
85 - 112	Reveresec order of the reference and processed speech samples in cond. 57 - 84			
Notes:	- 4 talkers are used for all conditions: 2 male and 2 female - 12 speech samples (4 s) are used for each talker - AMR@12.2 means AMR at 12.2 kbit/s - AMR/NS@12.2 means NS candidate x with AMR at 12.2 kbit/s			

Table 1.2: **Test conditions for Experiment 1**

II. Experiments 2a, 2b & 2c: Artifacts and Clipping Effects in Background Noise Conditions (DCR)

II.1. Introduction

These experiments were prepared to test requirements 4.5.1.3 and 4.5.1.4 in the associated Section in Stage 1 Description (TS GSM 02.76 v. 2.0.0), i.e. **No artifacts in residual noise, and No speech clipping and no reduction in intelligibility**.

Within standardisation bodies (ITU-T and ETSI), several procedures have been designed in the past to evaluate the effect of environmental noise, among then the ACR method using the classical Quality scale and the ACR method using the Listening Effort scale which were found to fail to prevent the noise from being the predominant factor within the test. During the TCH-HS (HR) Characterisation Phase, two methodologies : ACR (Quality scale) and DCR (Degradation scale) methods were followed to formally compare the two disting methods with exactly the same experimental design test plan. As the end of the exercise, it was proved that the DCR mode of collecting the subject's responses is the most appropriate for this purpose : it is a more discriminant procedure than the ACR method and it allows to draw better conclusions on the real quality performance of telecommunication systems when the input has been corrupted by background noise. This procedure is now known as the Modified DCR method and is widely used for assessing the quality performance of candidate codecs and standards with environemental noise.

The method of assessment which is proposed for all experiments #2 will, then, be the modified version of the Degradation Category Rating method (DCR, [2] where a quality (unprocessed, but noisy) reference is introduced prior to each evaluation). The instructions which will be given to the subjects are slithly modified (see Annex) in order to allow a possible answer : Degradation not perceived or even some improvement (score : 5), this modification was already introduced within past test plans, ITU-T Wideband (7 kHz) Selection for example [4]. These DCR experiments will be run for three types of acoustic background noise. The speech material used in these experiments are 8s sentence pair samples.

II.2 Test Factors and Conditions

The DCR test will be run for the following three types of acoustic background noise:

- A car noise that is stationary both in level and in spectrum.
- A street noise that is non-stationary in level but fairly stationary in spectrum.
- A babble noise that is fairly stationary in level but non-stationary in spectrum.

This results in a total of three DCR experiments with the different noise types in separate experiments. Within each experiment, a low and a high SNR level will be tested. The values for the low SNR are SNR_C = 6 dB for the car noise, SNR_S = 9 dB for the street noise, and SNR_B = 9 dB for the babble noise. The higher SNR will be equal to SNR + 6 dB for all three noise types. The noise samples will have been recorded in scenarios representative of the respective low SNR value for each noise type (i.e. SNR = 6 or 9 dB).

A coder tandem condition is included with each experiment. This condition is tested for the low SNR level only.

Main Codec Conditions	***#***	***Notes***
Noise Suppresser Candidate	1	
Codec	1	AMR
Codec Modes (FR/HR)	FR	12.2 kbps rate
	HR	5.9 kbps rate
BERs	0	Clear channel, no transmission errors
Input level	1	nominal: -26dB relative to OVL
Acoustic Background Noise	3	Static Car @ 6dB and 12dB
		Street @ 9dB and 15dB
		Babble @ 9dB and 15dB
Tandeming	1	Self tandem condition at one noise level
Input Characteristic	1	GSM Filtered
Codec references	***#***	***Notes***
All Experiments	1	AMR wo/ NS
Other references	***#***	***Notes***
Direct		nominal level, GSM Filtered
MNRU, Exp 2a, 2b, 2c		nominal level, with background noise, GSM Filtered, Q= 6, 12, 18, 24, 30dB
Ideal Noise Suppression	9	5 levels from SNR+6 for AMR at 12.2 kbps, 4 levels from SNR for AMR at 5.9 kbps
Common Conditions	***#***	***Notes***
GSM Channel	0	NO channel model
Number of talkers	4	2 male + 2 female
Number of speech samples	28	6/ talker for the main test + 1/ talker for the Practice session
Listening Level	1	-15dBPa (79dB SPL) at ERP
Listeners	24	Naive Listeners
Randomizations	6	6 groups of 4 listeners
Rating Scale	1	Modified DCR Instructions
Replications	1	Original Presentation Only

Table 2.1**: Factors and conditions for Experiments 2a, 2b, 2c**

II.3 Test Conditions for Experiment 2

Cond.	SNR value	Ideal NS (dB)	Trans-codings	Quality Ref. (A)	Codec. (B)	Speech sample number (6 sequences)
1	SNR	-	-	DIRECT	DIRECT	4 3 4 5 6 1
2	-	-	-	DIRECT	MNRU-30	4 3 4 5 6 1
3	-	-	-	DIRECT	MNRU-24	4 3 4 5 6 1
4	-	-	-	DIRECT	MNRU-18	4 3 4 5 6 1
5	-	-	-	DIRECT	MNRU-12	4 3 4 5 6 1
6	-	-	-	DIRECT	MNRU-6	4 3 4 5 6 1
7	SNR+6	-	1	DIRECT	AMR@12.2	1 2 3 4 5 6
8	SNR+6	4	1	DIRECT	AMR@12.2	1 2 3 4 5 6
9	SNR+6	6	1	DIRECT	AMR@12.2	1 2 3 4 5 6
10	SNR+6	8	1	DIRECT	AMR@12.2	1 2 3 4 5 6
11	SNR+6	10	1	DIRECT	AMR@12.2	1 2 3 4 5 6
12	SNR+6	12	1	DIRECT	AMR@12.2	1 2 3 4 5 6
13	SNR	-	1	DIRECT	AMR@5.9	2 3 4 5 6 1
14	SNR	4	1	DIRECT	AMR@5.9	2 3 4 5 6 1
15	SNR	6	1	DIRECT	AMR@5.9	2 3 4 5 6 1
16	SNR	8	1	DIRECT	AMR@5.9	2 3 4 5 6 1
17	SNR	10	1	DIRECT	AMR@5.9	2 3 4 5 6 1
18	SNR	-	2	DIRECT	AMR@12.2	3 4 5 6 1 2
19	SNR	-	1	DIRECT	AMR/NS@12.2	2 3 4 5 6 1
20	SNR	-	1	DIRECT	AMR/NS@10.2	2 3 4 5 6 1
21	SNR	-	1	DIRECT	AMR/NS@7.4	2 3 4 5 6 1
22	SNR	-	1	DIRECT	AMR/NS@6.7	2 3 4 5 6 1
23	SNR	-	1	DIRECT	AMR/NS@5.9	2 3 4 5 6 1
24	SNR	-	1	DIRECT	AMR/NS@5.15	2 3 4 5 6 1
25	SNR+6	-	1	DIRECT	AMR/NS@12.2	1 2 3 4 5 6
26	SNR+6	-	1	DIRECT	AMR/NS@10.2	1 2 3 4 5 6
27	SNR+6	-	1	DIRECT	AMR/NS@7.4	1 2 3 4 5 6
28	SNR+6	-	1	DIRECT	AMR/NS@6.7	1 2 3 4 5 6
29	SNR+6	-	1	DIRECT	AMR/NS@5.9	1 2 3 4 5 6
30	SNR+6	-	1	DIRECT	AMR/NS@5.15	1 2 3 4 5 6
31	SNR	-	2	DIRECT	AMR/NS@12.2	3 4 5 6 1 2
32	SNR	-	2	DIRECT	AMR/NS@10.2	3 4 5 6 1 2
33	SNR	-	2	DIRECT	AMR/NS@7.4	3 4 5 6 1 2
34	SNR	-	2	DIRECT	AMR/NS@6.7	3 4 5 6 1 2
35	SNR	-	2	DIRECT	AMR/NS@5.9	3 4 5 6 1 2
36	SNR	-	2	DIRECT	AMR/NS@5.15	3 4 5 6 1 2

Notes:
- 4 talkers are used for all conditions: 2 male and 2 female
- 6 speech samples (16 s) are used for each talker
- AMR@12.2 means AMR at 12.2 kbit/s
- AMR/NS@12.2 means NS candidate with AMR at 12.2 kbit/s

Experiment 2a: Car noise with SNR = SNR_C = 6 dB,
Experiment 2b: Street noise with SNR = SNR_S = 9 dB
Experiment 2c: Babble noise with SNR = SNR_B = 9 dB

Table 2.2 **: Test conditions for Experiments 2a, 2b, 2c**

REFERENCES

[1] **Test Plan Specification for the AMR-NS Selection Phase,** Tdoc SMG11/S4 356/99 R3, Source: AMR-NS/SQ

[2] **ITU-T Rec. P.800 "Methods for subjective determination of transmission quality"**

[3] **ITU-T Com 12 : Hanbook on Telephonometry**

[4] **Subjective Selection Test Plan for the ITU-T Wideband (7 kHz) Speech Coding Algorithm,** Version 1.2, May 1998

ANNEX

Table A.1 - Example of instructions to subjects for DCR and Modfied-DCR test.

INSTRUCTIONS TO SUBJECTS
(Modified) Degradation category rating test
"Evaluation of the influence of various environmental noises on the quality of different telephone systems"

You are going to hear, through the headphones which is in front of you, various pairs of speech samples recorded in different noise environments (for example inside a car or an office).

Each pair is build up from two samples separated by a pause of about half a second.

Each sample is build up from two speech sentences separated by a small gap. All the samples have been recorded in the different noise environments.

Within each pair of samples, the first is the reference and the second one, which has been treated by telephone system, has to be evaluated with regard to the reference.

You are kindly requested to listen carefully to each pair of samples. Then when the green light is on, please record your opinion about the modifications perceived on the second sample with regard to the first one (reference) using the following scale :

5	Degradation not perceived **or even some improvement**
4	Degradation perceived but not annoying
3	Degradation slightly annoying
2	Degradation annoying
1	Degradation very annoying

You will have 5 seconds to record your answer by pushing the button corresponding to your choice.

Then you will have a short pause before the presentation of next pair.

We will start by a short practice session to make you familiar with the test procedure. Then the actual tests will take place during sessions of 10 to 15 minutes.

Listener name : Date :

Group number : Table n° :