ETSI SMG11 **Tdoc SMG11/NS 420/99**

Kyoto, Japan 6-10 December 1999

Title: **ETSI/AMR NS Selection Experiment 2 - COMSAT Results**

Source*: **COMSAT Laboratories**

Summary

This document presents a summary for ETSI Adaptive Multi Rate (AMR) Noise Selection (NS) Selection Phase Experiment 2 conducted by COMSAT in the French language.

1. Introduction

COMSAT Laboratories performed a listening laboratory evaluation in French in accordance with the AMR NS Selection Phase Experiment 2. This experiment was designed to assess whether the noise suppression candidate algorithms for the AMR codec operating at 12.2 and 5.9 kbit/s would introduce any degradation for a quiet background for one and two transcodings. The test design is defined in Section 8 of the AMR NS Selection Subjective Test Plan [1]. COMSAT performed Experiment 2 using a subset of the French speech material available in the NTT Speech Database. Forty-eight native speakers of the French language (instead of 24, due to listening session duration constraints) performed as subjects in the test, which was nominally balanced for gender. The raw data collected was used to derive gender-wise and combined-gender pair comparison scores and standard deviation statistics.

2. Source Material

Thirteen single-sentence stimuli were selected for two male and two female talkers from the NTT Speech Database, for a total of 52 different source speech stimuli. Twelve single-sentences were allocated per talker for the main assessment sessions, and one single-sentence was allocated per talker for the practice session. All files had an exact duration of 4 seconds. The source material was provided to COMSAT, the designated Host Laboratory, which was responsible for all pre- and post-processing according to [2].

3. Experimental Design

The test design followed the specification in the AMR NS Selection Test Plan, as summarized in Tables 1 and 2.

4. Processed Material

The host laboratory provided a CDROM with 2720 processed speech files, which corresponds to the processing of (12+1) single-sentence per talker for four talkers through 28 test conditions. See [3] for details on the source speech processing.

Contact: Simão F. Campos-Neto
COMSAT Labs
22300 COMSAT Drive
Clarksburg, MD 20871
USA

Tel: +1-301-428-4516
Fax: +1-301-428-9287
E-mail: simao.campos@comsat.com

Table 1:
Factors and Conditions for Experiment 2.

Main Codec Conditions	**#**	**Notes**
Noise Suppresser Candidates	6	
Codec	1	AMR
Codec Modes (FR/HR)	HR	5.9 kbit/s rate
	FR	12.2 kbit/s rate
BER	0	Clear channel, no transmission errors
Input level	1	Nominal (-26dBov)
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	2	AMR@12.2, AMR@5.9
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	48	Naive Listeners
Randomizations	6	6 groups of 4 listeners
Rating Scale	1	PC Instructions
Replications	2	Original presentation order for the first 24 subjects and the modified presentation order for the subsequent 24 subjects.

5. Listening Sessions

COMSAT performed Experiment 2 according to the Test Plan, with the exception that instead of 24 subjects listening to two replications of the processed material, twice as many subjects were used. This was done because the procedure defined in the Test Plan would result in sessions longer than two hours, which COMSAT did not deem appropriate. The option to summon the same subjects at a later date was not a viable solution, in particular because of the tight time schedule proposed for the activity. The most appropriate solution that would still accomplish the objectives of the test design was to use twice the number of subjects, as described below.

5.1 Presentation Sequence Material

COMSAT used the grouping and randomization sequences specified in the AMR NS Selection Test Plan for Experiment 2.

Table 2:
Experiment 2: Degradation in Clean Speech

Cond.	Reference Codec	Processed Codec	Trans-codings
1	AMR@12.2	AMR@12.2	1
2	AMR@12.2	AMR@12.2	2
3	AMR@5.9	AMR@5.9	1
4	AMR@5.9	AMR@5.9	2
5	AMR@12.2	AMR/NS1@12.2	1
6	AMR@12.2	AMR/NS2@12.2	1
7	AMR@12.2	AMR/NS3@12.2	1
8	AMR@12.2	AMR/NS4@12.2	1
9	AMR@12.2	AMR/NS5@12.2	1
10	AMR@12.2	AMR/NS6@12.2	1
11	AMR@12.2	AMR/NS1@12.2	2
12	AMR@12.2	AMR/NS2@12.2	2
13	AMR@12.2	AMR/NS3@12.2	2
14	AMR@12.2	AMR/NS4@12.2	2
15	AMR@12.2	AMR/NS5@12.2	2
16	AMR@12.2	AMR/NS6@12.2	2
17	AMR@5.9	AMR/NS1@5.9	1
18	AMR@5.9	AMR/NS2@5.9	1
19	AMR@5.9	AMR/NS3@5.9	1
20	AMR@5.9	AMR/NS4@5.9	1
21	AMR@5.9	AMR/NS5@5.9	1
22	AMR@5.9	AMR/NS6@5.9	1
23	AMR@5.9	AMR/NS1@5.9	2
24	AMR@5.9	AMR/NS2@5.9	2
25	AMR@5.9	AMR/NS3@5.9	2
26	AMR@5.9	AMR/NS4@5.9	2
27	AMR@5.9	AMR/NS5@5.9	2
28	AMR@5.9	AMR/NS6@5.9	2

5.2 Listeners

The subjective assessment was performed using 48 listeners (nominally balanced between male and female), divided in two rounds of six groups of four listeners each. The first 24 subjects listened to the presentation order defined in Annex F.4 of the Test Plan, and the subsequent 24 subjects listened to the presentation order modified as per the instructions in Annex F.4.

The listener selection criteria were compliant with the AMR NS Selection Test Plan, noting that Audiometric testing was not performed on the listeners, for legal reasons. Test subjects were selected from an existing pool of native French language listeners for which past assessment performance data exists indicating their general hearing integrity. Subject performance within this experiment was compared to the overall performance of all listeners used in each experiment as a check on the hearing integrity of each listener at the time of testing. The pre-test listener orientation used by COMSAT conformed to that specified in the Test Plan.

5.3 Audio Presentation

The processed speech material was presented to groups of listeners, seated at separate, visually screened listening stations contained within an acoustically conditioned sound room meeting the requirements for an NC 20 acoustic facility. Presentation was made monaurally using a telephone handset, driven by a distribution amplifier set to deliver monophonic speech to the listener's preferred listening ear at an active level of -15 dBPa (79 dB SPL), using a B&K 4153 Artificial Ear with circumaural headphone adapter, 4134 Microphone element and 2610 Measurement Amplifier.

The processed speech files were stored within the main facility computer and presented to the listeners under program control as 16 kHz samples through a 16-bit, D/A coupled to the input of the distribution amplifier through a Frequency Devices 9002 Eight-pole Elliptic Filter, set for a bandpass of 200-Hz to 3.4-kHz. Auxiliary filtering was performed to achieve an overall modified-IRS receive characteristic.

The listener responses were registered on auxiliary computers. One of these voting terminals is contained within each voting station. Voting was permitted following the completed presentation of each voting stimulus (in this experiment, a pair of associated single-sentences). All seated listeners were required to register responses prior to the subsequent presentation of a new stimulus. Once a group of listeners was conditioned to the dynamics of the voting procedure, the voting response time for each presented stimulus was nominally three seconds for each presented stimulus.

5.4 Scoring

Within experiments using a pair comparison method of assessment, the presented single-sentences were scored by the listeners using a pair comparison scale either as the first stimulus being preferred over the second (encoded by the computer as a vote 1), or otherwise (encoded by the computer as a vote 0). Subjects were forced to choose one of the options, as equal-preference votes are not allowed in standard pair comparison tests. The semantic designations were presented on the screen of the voting terminals and selected through the use of a pointing device. The voting screen was rendered neutral during the presentation of each new stimulus.

As all seated listeners completed their voting, the votes for all stations were transferred to the main facility computer prior to the presentation of subsequent new material. The votes of each group of listeners for each presentation set of speech material were stored as ASCII files within the main facility computer for subsequent analysis and presentation.

Upon completion of the listening sessions, all raw data were de-scrambled and consolidated into a single ASCII file, which was used for the statistical analysis.

6. Statistical Analysis

The statistics to be reported for this pair-comparison experiment [4] are the proportion *P* of subjects preferring the test stimulus over the reference stimulus (as defined in Table 2) for a total of *N* votes per condition, the standard deviation *s*:

$$s = \sqrt{\frac{P \cdot (1-P)}{N}} \qquad \textbf{(Eq.1)}$$

and the upper and lower confidence limits, as calculated by:

$$CI_{1-\alpha} = \frac{N}{N + z_{1-\alpha/2}^2} \cdot \left(P + \frac{z_{1-\alpha/2}^2}{2N} \pm z_{1-\alpha/2} \sqrt{\frac{P \cdot (1-P)}{N} + \frac{z_{1-\alpha/2}^2}{4N^2}} \right) \qquad \textbf{(Eq.2)}$$

where $z_{1-\alpha/2}$ is the standardized score for a normal distribution cutting off the lower $\alpha/2$ proportion of cases.

Additionally, a hypothesis to test was whether the preference for the noise reduction-enabled AMR codec was statistically different from the ideal proportion π=0.5, i.e. that the AMR with noise suppression is equally preferred to AMR without noise suppression (for quiet background). In other words,

$$H_0: \quad \pi = 0.5$$
$$H_1: \quad \pi \neq 0.5$$

The null hypothesis Ho is tested using a *z* test where:

$$z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{N}}} = \frac{P - 0.5}{0.5/\sqrt{384}} = 2\sqrt{384} \cdot (P - 0.5) = 39.192 \cdot (P - 0.5) \qquad \textbf{(Eq.3)}$$

Hence, the null hypothesis is rejected if

$$|z| \geq z_{1-\alpha/2}$$

Or accepted if:

$$0.5 - \frac{z_{1-\alpha/2}}{39.19} < P < 0.5 + \frac{z_{1-\alpha/2}}{39.19} \qquad \textbf{(Eq.4)}$$

For a 95% confidence level, Equations 2 and 4 are reduced to ($z_{1-\alpha/2} = 1.96$, N=384):

$$CI_{95\%} = \frac{N}{N+3.84}\left[P + \frac{1.92}{N} \pm 1.96\sqrt{\frac{P\cdot(1-P)}{N} + \frac{0.96}{N^2}}\right] \approx 0.99\left[P + 0.005 \pm 0.1\sqrt{P(1-P)}\right] \qquad \textbf{(Eq.5)}$$

$$0.45 < P < 0.55 \qquad \textbf{(Eq.6)}$$

Table 3 presents the basic statistical analysis data produced by COMSAT for AMR NS Selection Experiment 2, similar to the data provided to the Global Analysis Laboratory. Each test condition received a total of 96 votes. In the table, *Cnd* represents the test condition number, *P* is the proportion of subjects preferring the Test Codec stimulus, *s* is the standard deviation, +95% and -95% represent the upper and lower (1-α)=95% confidence limits (as calculated by Eq.5), and the *Ho* column identifies whether the null hypothesis is accepted (A) or rejected (R), as calculated by Eq.6.

It can be seen from Table 3 that, at the 95% confidence level, NS 1 failed the null hypothesis for one transcoding at 12.2 kbit/s, NS 2 failed the null hypothesis for two transcodings at 12.2 kbit/s, and NS 4 failed the null hypothesis for both 12.2 kbit/s and 5.9 kbit/s two-transcoding cases.

7. Conclusion

COMSAT performed AMR NS Selection Experiment 2 for the French language in compliance with the test plan, with the exception that 48 subjects were used, instead of 24, to achieve the desired test power while keeping the listening assessment sessions within the time limit usually used by COMSAT. This experiment was designed to assess whether the noise suppression candidate algorithms for the AMR codec operating at 12.2 and 5.9 kbit/s would introduce any degradation for a quiet background. It was observed that the basic hypothesis that the subjects would equally prefer the AMR codec with noise suppression over the AMR codec without noise suppression in a quiet background was not valid at the 95% confidence level for NS candidate 1 for one transcoding at 12.2 kbit/s, for NS candidate 2 for two transcodings at 12.2 kbit/s, and for NS candidate 4 at 12.2 kbit/s and 5.9 kbit/s in the two-transcoding case. In all other cases, the subjects equally preferred the AMR codec with or without noise suppression for a quiet background.

References

[1] SMG11SQ, *"Test Plan for the AMR Specification for the AMR-NS Selection Phase"*; Tdoc SMG11 288/99.

[2] SMG11SQ, *"Processing Functions for the GSM AMR Noise Suppression Selection Tests"*; Tdoc SMG11 281R/99

[3] COMSAT Laboratories, *"Host Laboratory Processing for ETSI/AMR Noise Suppression Selection Tests"*, Tdoc SMG11 417/99

[4] Glass, G.V; Hopkins, K.D., "Statistical Methods in Education and Psychology"; 3rd edition; Allyn & Bacon, Needham Heights, MA 02194, 1995; pp.319-330.

Table 3

MOS, standard deviation, and 95% confidence interval for COMSAT's Experiment 2

Cnd	Reference (Reference Codec)	Processed (Test Codec)	Tandem	N	P	s	+95%	-95%	H_0 ?
1	AMR@12.2	AMR@12.2	1	384	0.4661	0.0255	0.5158	0.4170	A
2	AMR@12.2	AMR@12.2	2	384	0.5000	0.0255	0.5495	0.4505	A
3	AMR@5.9	AMR@5.9	1	384	0.5260	0.0255	0.5752	0.4763	A
4	AMR@5.9	AMR@5.9	2	384	0.5234	0.0255	0.5726	0.4737	A
5	AMR@12.2	AMR/NS1@12.2	1	384	0.4167	0.0252	0.4663	0.3686	**R**
6	AMR@12.2	AMR/NS2@12.2	1	384	0.4688	0.0255	0.5184	0.4196	A
7	AMR@12.2	AMR/NS3@12.2	1	384	0.4974	0.0255	0.5469	0.4479	A
8	AMR@12.2	AMR/NS4@12.2	1	384	0.5052	0.0255	0.5546	0.4556	A
9	AMR@12.2	AMR/NS5@12.2	1	384	0.4922	0.0255	0.5417	0.4427	A
10	AMR@12.2	AMR/NS6@12.2	1	384	0.5000	0.0255	0.5495	0.4505	A
11	AMR@12.2	AMR/NS1@12.2	2	384	0.5130	0.0255	0.5623	0.4634	A
12	AMR@12.2	AMR/NS2@12.2	2	384	0.4115	0.0251	0.4610	0.3636	**R**
13	AMR@12.2	AMR/NS3@12.2	2	384	0.5026	0.0255	0.5520	0.4530	A
14	AMR@12.2	AMR/NS4@12.2	2	384	0.4349	0.0253	0.4846	0.3864	**R**
15	AMR@12.2	AMR/NS5@12.2	2	384	0.5156	0.0255	0.5649	0.4659	A
16	AMR@12.2	AMR/NS6@12.2	2	384	0.4635	0.0254	0.5132	0.4145	A
17	AMR@5.9	AMR/NS1@5.9	1	384	0.4870	0.0255	0.5365	0.4376	A
18	AMR@5.9	AMR/NS2@5.9	1	384	0.4948	0.0255	0.5443	0.4453	A
19	AMR@5.9	AMR/NS3@5.9	1	384	0.5052	0.0255	0.5546	0.4556	A
20	AMR@5.9	AMR/NS4@5.9	1	384	0.4922	0.0255	0.5417	0.4427	A
21	AMR@5.9	AMR/NS5@5.9	1	384	0.5182	0.0255	0.5675	0.4685	A
22	AMR@5.9	AMR/NS6@5.9	1	384	0.4557	0.0254	0.5054	0.4068	A
23	AMR@5.9	AMR/NS1@5.9	2	384	0.5026	0.0255	0.5520	0.4530	A
24	AMR@5.9	AMR/NS2@5.9	2	384	0.4948	0.0255	0.5443	0.4453	A
25	AMR@5.9	AMR/NS3@5.9	2	384	0.5026	0.0255	0.5520	0.4530	A
26	AMR@5.9	AMR/NS4@5.9	2	384	0.4427	0.0253	0.4924	0.3941	**R**
27	AMR@5.9	AMR/NS5@5.9	2	384	0.5156	0.0255	0.5649	0.4659	A
28	AMR@5.9	AMR/NS6@5.9	2	384	0.5156	0.0255	0.5649	0.4659	A