3GPP TR 06.78 V8.0.1 (2003-03)

Technical Report

3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Digital cellular telecommunication system (Phase 2+); Results of the AMR Noise Suppression Selection Phase (Release 1999)

Keywords
AMR, CODEC

3GPP

Postal address

3GPP support office address
650 Route des Lucioles - Sophia Antipolis
Valbonne - FRANCE
Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Internet
http://www.3gpp.org

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
  - 1 presented to TSG for information;
  - 2 presented to TSG for approval;
  - 3 or greater indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the document.

1 Scope

This technical report provides background information on the performance of the six candidates which were proposed as solutions for publication of an example noise suppression solution for application to the GSM Adaptive Multi-Rate (AMR) speech codec. Experimental test results from the speech quality related testing are reported to illustrate the behaviour of the candidate algorithms in multiple operational conditions. Additional information is also provided covering data not necessarily directly associated with speech quality (such as complexity, delay, effect on voice activity factor).

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] GSM 01.04: "Digital cellular telecommunications system (Phase 2+); Abbreviations and acronyms".

[2] GSM 02.76: "Noise Suppression for the AMR Codec; Service Description; Stage 1"

[3] GSM 03.50: "Transmission planning aspects of the speech service in the GSM Public Land Mobile Network (PLMN) system".

[4] GSM 06.08: "Digital cellular telecommunications system; Half rate speech; Performance of the GSM half rate speech codec".

[5] GSM 06.55: "Digital cellular telecommunications system; Performance Characterisation of the GSM Enhanced Full Rate (EFR) speech codec".

[6] GSM 06.75: "Digital cellular telecommunications system; ; Performance Characterisation of the GSM Adaptive Multi-Rate (AMR) speech codec".

3 Definitions and abbreviations

3.1 Definitions

The following terminology is used throughout this report.

Adaptive Multi-Rate (AMR) codec; Speech and channel codec capable of operating at gross bit-rates of 11.4 kbit/s ("*half-rate*") and 22.8 kbit/s ("*full-rate*"). In addition, the codec may operate at various combinations of speech and channel coding (*codec mode*) bit-rates for each *channel mode*.

Channel mode; Half-rate or full-rate operation

Codec mode; For a given *channel mode*, the bit partitioning between the speech and channel codecs.

Error Patterns

Error Insertion Device; Result of offline simulations stored on files. To be used by the "Error Insertion Device" to model the radio transmission from the output of the channel decoder and interleaver to the input of the deinterleaver and channel decoder.

Full-rate (FR); Full-rate channel or *channel mode*

Half-rate (HR); Half-rate channel or *channel mode*

Toll Quality; Speech quality normally achieved on modern wireline telephones. Synonymous with "ISDN quality".

Wireline quality; Speech quality provided by modern wireline networks. Normally taken to imply quality at least as good as that of 32kbit/s G.726 or G.728 16 kbit/s codecs.

3.2 Abbreviations

For the purposes of the present document, the following abbreviations apply:

A/D	Analogue to Digital
ACR	Absolute Category Rating
ADPCM	Adaptive Differential Pulse Code Modulation
AMR	Adaptive Multi-Rate
AMR-NS	AMR Noise Suppression
BSC	Base Station Controller
BTS	Base Transceiver Station
CCR	Comparison Category Rating
C/I	Carrier-to-Interfere ratio
CI	Confidence Interval
CNI	Comfort Noise Insertion
CRC	Cyclic Redundancy Check
D/A	Digital to Analogue
DAT	Digital Audio Tape
DCR	Degradation Category Rating
DSP	Digital Signal Processor
DTMF	Dual Tone Multi Frequency
DTX	Discontinuous Transmission for power consumption and interference reduction
EFR	Enhanced Full Rate
ESP	Product of E (Efficiency), S (Speed) and P (Percentage of Power) of the DSP
FR	Full Rate (also GSM FR)
FH	Frequency Hopping
G.726	ITU 16/24/32kbit/s ADPCM codec
G.728	ITU 16kbit/s LD-CELP codec
G.729	ITU 8/6.4/11.8 kbit/s speech codec
GBER	Average gross bit error rate
GSM	Global System for Mobile communications
HR	Half Rate (also GSM HR)
IRS	Intermediate Reference System
ITU-T	International Telecommunication Union - Telecommunications Standardisation Sector
MNRU	Modulated Noise Reference Unit
Mod. IRS	Modified IRS
MOPS	Million of Operation per Seconds
MOS	Mean Opinion Score
MS	Mobile Station
MSC	Mobile Switching Center
PCM	Pulse Code Modulation
PSTN	Public Switched Telecommunications Network
Q	Speech-to-speech correlated noise power ratio in dB
SD	Standard Deviation
SID	Silence Descriptor
SMG	Special Mobile Group
SNR	Signal To Noise Ratio
TCH-AFS	Traffic CHannel Adaptive Full rate Speech
TCH-AHS	Traffic CHannel Adaptive Half rate Speech

TDMA	Time Division Multiple Access
TFO	Tandem Free Operation
tMOPS	true Million of Operations per Seconds
TUx	Typical Urban at multipath propagation profile at x km/s
VAD	Voice Activity Detector
VAF	Voice Activity Factor
wMOPS	weighted Million of Operations per Seconds

Multiple Error Patterns were used during the Characterisation tests. They are identified by the propagation Error Conditions from which they are derived. The following conventions are used:

ECx	Error Conditions at x dB C/I simulating a radio channel under static C/I using ideal Frequency Hopping in a TU3 multipath propagation profile

For abbreviations not given in this sub-clause, see GSM 01.04 [1].

4 General

4.1 Project History

In June 1998 during SMG#26, SMG approved a Work Item to develop and standardise a noise suppression solution for the Adaptive Multi-rate (AMR) speech codec. SMG11 have carried out this work since September 1998 (SMG11#7).

The work in SMG11 focussed on the preparations for a Selection Phase with the intention of choosing an example optional noise suppression solution. In the course of this work, documentation covering Requirements [2], Design Constraints, Selection Phase Deliverables, Selection Phase Rules, and a Selection Phase Test Plan were drafted.

In August 1999 the Selection Phase commenced. Six Noise Suppression algorithms were submitted as candidates. The algorithm proposals came from Ericsson (NS5), Matra Nortel Communications (MNC) (NS4), Mitsubishi Electric Corporation (NS1), Motorola (NS6), Nokia (NS3) and Siemens AG (NS2). Testing of candidate solutions was carried out during September-November 1999, and the listening test results were analysed at two meetings: SMG11#13 (December 1999) and SMG11#14 (January 2000). Listening test results and deliverables from proponents (technical descriptions of the algorithms, analysis of compliance to design constraints, additional information such as objective measurements) were reviewed within SMG11.

SMG11 were not able to reach a consensus on selecting an example solution, and as a consequence the deliverables from the Work Item were amended during SMG#31 to comprise of a specification defining Recommended Minimum Performance Requirements [TBA], an associated Subjective Listening Test Plan, and a Technical Report recording all pertinent information arising from the Selection Phase. This document forms the latter deliverable.

4.2 Overview of the AMR-NS Work Item

The Work Item covered the development of a noise suppression algorithm as an example optional feature designed to enhance speech quality in a range of environments where there is significant (acoustic) background noise. The noise suppression function is a preprocessing module that is used to improve the signal to noise ratio of a speech signal prior to voice coding. Solutions implementing noise suppression as a separate preprocessing module prior to the AMR speech encoder or as an embedded module operating on the input speech buffer were considered. AMR Noise Suppression (AMR-NS) is intended to be used in the mobile station (operating on the uplink speech signal). The possibility to implement AMR Noise Suppression in the network (operating on the downlink speech signal) was considered for feasibility purposes only. As part of this study, tests with noise suppression in both uplink and downlink (tandem noise suppression) were included and the results are included in this report. It should be noted that the Recommended Minimum Performance Requirements Specification [TBA] covers only the uplink case where the algorithm is implemented in the mobile station.

4.3 Presentation of the following sections

The following sections provide a summary of the Selection Phase test results, including the results of objective performance measurements, and a record of relevant other information for each of the candidate algorithms.

- Section 5 defines the minimum performance requirements defined for the Selection Phase.
- Section 6 defines the means used to compare candidate algorithms directly in terms of speech quality performance.
- Section 7 describes the subjective listening tests undertaken and summarises the results achieved (covering the requirements of Section 5 and the means of comparison of Section 6).
- Section 8 summarises the design constraints defined for the Selection Phase.
- Section 9 summarises the effect on the existing AMR Voice Activity Detector (VAD) function, in the form of voice activity factor (VAF) measurements.
- Section 10 summarises the results of an Objective Performance Measure used to characterise the noise suppression algorithms.
- Section 11 summarises the results of the Feasibility study into Implementing Noise Suppression in the downlink.

Annex A contains the final versions of the Design Constraints, Selection Rules, and Selection Phase Deliverables defined for the Selection Phase

Annex B (a separate component of the archive file comprising this report) is the final version of the Selection Phase Test Plan.

Annex C (a separate component of the archive file comprising this report) is the final version of the Selection Phase Global Analysis Spreadsheet, and is the full record of the results achieved from the subjective listening tests.

Annex D contains the methodologies used to derive signal to noise ratio improvement values from the subjective listening tests.

Annex E defines the methodology used to generate the objective measures of performance reported in section 10.

Annex F defines the methodology used to determine impact on Voice Activity Factor.

Annex G provides a reference list of SMG11 temporary documents which contain relevant information used during the Selection Phase. This includes references to the final versions of the reports provided by the listening laboratories.

5 Minimum Performance Requirements

Performance requirements were established during the AMR-NS development phase which reflected the understanding that there was no clear risk-free means of identifying minimum performance, because this was the first time such a standardisation effort for noise suppression functionality had been undertaken in ETSI. As a result, failure to meet some minimum performance requirements was not considered to be a reason for disqualification, particularly if such failures were not consistent across all listening laboratories undertaking a particular test (where the term systematic failure is used to describe failures consistent across laboratories).

Table 5.1 lists the minimum requirements as stated in the Stage 1 specification [2] and, for each requirement, defines the associated experiment or experiments defined to check compliance to the requirement. In each case a criterion is defined to determine failure to meet the requirement. The reference condition is AMR without noise suppression in all cases, except for the evaluation of speech quality during the Initial Convergence Time (Experiment 1).

The possibility to implement AMR Noise Suppression in the network (operating on the downlink speech signal) was defined to be part of a feasibility study. This is considered further in Section 11.

Associated Section in Stage 1 Description [2]	Requirement (Title)	Relevant Tests
4.6.1.1	Initial Convergence	Experiment 1: Expert/Informal listening test Any candidate for which the Listening Experts determine that the quality degradation in the initial convergence time is unacceptable will be regarded as failing the requirement.
4.6.1.2	No degradation in clean speech	Experiment 2: Degradation in Clean (Pair comparison) Any candidate failing to be preferred with a 50% probability in any test condition will be regarded as failing the requirement
4.6.1.3/4.6.1.4	No artefacts in residual noise & No speech clipping or reduction in intelligibility	Experiment 3: Performance in Background Noise (ACR) A candidate failing to be at least as good as AMR without NS at the same noise level will be regarded as failing the requirement
4.6.1.5	AMR+NS preferred to AMR without NS	Experiments 4-10: Performance under background noise. Any candidate failing to be preferred to the reference (AMR without NS) with a 95% probability for any condition will be regarded as failing the requirement.
4.8	Voice Activity Factor	Test defined in [3] section 4.8:Any candidate failing to meet the requirement stated in section 4.8 of [2] will be regarded as failing the requirement. (This requirement states that the use of noise suppression should not significantly increase channel activity when used in conjunction with DTX.)

Table 5.1: *Minimum performance requirements*

The total number of simple failures and number of systematic failures (failure of the same test condition in all tests performed for the same experiment) were recorded and the candidates were ranked accordingly.

In order to generate additional information, the candidates were also ranked according to the number of simple and systematic failures, assuming that a candidate only fails as a result of Experiments 4-10 if it is not found at least as good as the reference at the 95% confidence interval.

6 Comparison of Candidates by Subjective Means

A number of means of ranking candidates were developed based on figures of merit (FOMs). These FOMs were derived from the listening test results and are defined below. The FOMs covering downlink operation are not included (see Section 11), and FOMs defined in the Selection Rules which are not distinct are also not included. (It was originally intended to use weighted FOMs in addition to unweighted FOMs. Since no agreements were reached on weightings, the weighted FOMs are identical to the unweighted FOMs, and are therefore not reported here.) Additionally, FOMs not associated with subjective listening test results are not included.

Two sets of Figure of Merits are described here. The FOM numbering definitions used during the Selection Phase are retained for ease of cross-referencing. The first set (FoMs#1, 3, 4, 6) is based on the CCR test results (AMR/NS Selection Experiments 4 to 9). The second set (FoMs#7, 10) is derived from the ACR Test results.

FOM#1	Summation of CMOS scores for all conditions within an experiment, summed (unweighted) across all experiments, excluding all conditions including NS in the downlink direction.
	Repeat measures as per FOM#1above but restricted to:
FOM#3a	All conditions where the SNR >= 10dB (but not including conditions where the SNR >= 30dB)
FOM#3c	All conditions where the SNR < 10dB
	Repeat measures as per FOM#1 above but restricted to:
FOM#4a	All conditions with DTX on
FOM#4b	All conditions with DTX off
	Subjective SNR improvement per Noise Type based on the CCR test results evaluated using the methodology defined in Annex?:
FOM#6a	For car noise
FOM#6b	For street noise
FOM#6c	For babble noise
FOM#7a	Summation of the delta MOS scores per Experiment summed across all ACR Experiments (Experiment 3 and 10), excluding all test conditions using noise suppression in the downlink direction. The delta MOS score is identified as the difference between the MOS score obtained by the candidate for a specific test condition and the MOS score for the reference (AMR without noise suppression) in the same test condition.
FOM#7b	Unweighted summation of the delta dBq scores per Experiment summed across all ACR Experiments (Experiment 3 and 10), excluding all test conditions using noise suppression in the downlink direction. The delta dBq score is identified as the difference between the dBq score obtained by the candidate for a specific test condition and the dBq score for the reference (AMR without noise suppression) in the same test condition. When a dBq score is outside the linear part of the MNRU curve, it should be replaced by the dBq value obtained by replacing the non-linear part of the MNRU curve with a linear extrapolation with slope 0.05. The linear part of the MNRU curve is identified as the area of the *MOS=f(dBq)* curve where the slope is higher than 0.05.
FOM#10	Subjective SNR improvement per Noise Type based on the ACR test results evaluated using the methodology defined in Annex
FOM#10a	For car noise
FOM#10b	For street noise
FOM#10c	For babble noise

7 Selection Phase Listening Tests and Results

The candidates were referred to as NS1, NS2,..., NS6 during the analysis. The mapping to particular candidates is defined below.

NS1 = Mitsubishi Electric Corporation

NS2 = Siemens AG

NS3 = Nokia

NS4 = Matra Nortel Communications

NS5 = Ericsson

NS6 = Motorola

7.1 Summary of Selection Tests undertaken

The six candidates were tested in a variety of test conditions in 5 independent test laboratories. Testing was carried out using 6 languages. The tests took place during a period from September to November 1999.

Candidate performances were evaluated across many test conditions consisting of 10 experiments and 14 sub-experiments [Annex B]:

Experiment 1: *Quality During the Initial Convergence Time (informal test)*

Experiment 2:	*Degradation in Clean Speech (pair comparison test)*
Experiment 3:	*Artefacts and Clipping Effects in Background Noise Conditions (ACR test)* *Experiment 3a: car noise* *Experiment 3b: street noise* *Experiment 3c: babble noise*
Experiments 4 and 5:	*Performances in Background Noise Conditions (CCR test)* *Experiment 4a: low SNR, with AMR 5.9 kbit/s* *Experiment 4b: high SNR, with AMR 5.9 kbit/s* *Experiment 5a: low SNR, with AMR 12.2 kbit/s* *Experiment 5b: high SNR, with AMR 12.2 kbit/s*
Experiments 6 and 7:	*Performance in Background Noise: Influence of Propagation Errors (CCR test)* *Experiment 6: car noise at SNR of 6 dB with C/I=10 dB in uplink and error-free in downlink* *Experiment 7: street noise at SNR of 9 dB with C/I=10 dB in uplink and error-free in downlink*
Experiments 8 and 9:	*Performances in Background Noise: Influence of VAD/DTX (CCR test)*
Experiment 10:	*Influence of the Input Signal + Noise Level and Performances with Special Noises (ACR test)*

Experiment 1 is an informal test with expert listeners analysing any negative impact the noise suppressers may have during convergence time. *Experiment 2* is based on pair comparison to test if there is any degradation when using noise suppression compared to the coder without noise suppression. *Experiment 3* is an Absolute Category Rating (ACR) test analysing any artefacts and clipping effects in background noise. *Experiments 4 to 9* are Comparison Category Rating (CCR) tests analysing performances in background noise conditions with and without propagation errors, and also the influence of VAD/DTX. *Experiment 10* is an ACR test investigating the influence of the level of input signal and noise, and also assessing the performance for special noise types.

Most of the testing was carried out either as ACR or CCR tests. These two differ from each other in the methodology. ACR tests ask the listeners to assess the quality of each speech sample under test while CCR tests are based on asking the listeners to assess the quality differences between two samples. ACR and CCR tests are both well established and recognised speech quality testing methodologies.

The listening test laboratories performing the selection tests were: Arcon (English language), AT&T (Mandarin, Spanish and English), Nortel Networks (English), FUB (Italian), and COMSAT (French, Spanish and Japanese). All experiments and sub-experiments were carried out with 2 languages. The allocation of experiments to listening laboratories, and the languages used for each experiment, are shown in Table 7.1.

	Arcon	**AT&T**	**Nortel Networks**	**FUB**	**COMSAT**
	English	Mandarin Spanish, or English	English	Italian	French, Spanish, or Japanese
1			*X*		*Spanish*
2	*X*				*French*
3a	*X*			*X*	
3b	*X*			*X*	
3c	*X*			*X*	
4a			*X*		*Spanish*
4b			*X*		*Spanish*
5a			*X*		*Spanish*
5b			*X*		*Spanish*
6		*Spanish*	*X*		
7		*Spanish*	*X*		
8		*Mandarin, English*			
9		*Mandarin, English*			

10	*X*				*Japanese*
Host lab	ARCON	*COMSAT*	ARCON	*COMSAT*	*COMSAT*

Table 7.1: *Allocation of Experiments to Listening Laboratories*

The reference conditions were processed by Arcon and COMSAT, while the test samples were processed through the candidate algorithms by the candidate organisations themselves and were cross checked by other candidates. A blind procedure was followed to ensure that the test laboratories and the test subjects had no knowledge of the test conditions.

7.2 Summary of Listening Test Results Covering Minimum Performance Requirements

The candidates were ranked according to the number of simple and systematic failures (with the latter meaning failure of the same test condition in all tests performed for the same experiment).

All candidate algorithms failed to fulfil some of the minimum performance requirements. Table 7.2 records the number of failures and the ranking for each candidate according to the minimum performance requirements as stated in Table 5.1 (excluding those not associated with listening tests).

Simple Failures (excluding noise suppression in the downlink)	5	6	9	9	9	13
	1. NS5	**2. NS2**	**3. NS3**	**3. NS4**	**3. NS6**	**6. NS1**
Systematic Failures (excluding noise suppression in the downlink)	2	2	2	2	2	4
	1. NS2	**1. NS3**	**1. NS4**	**1. NS5**	**1. NS6**	**6. NS1**

Table 7.2: *Failures per candidate using the Minimum Performance Requirements*

Additionally results for the number of failures and the rankings are presented in Table 7.3 where the requirements of Table 5.1 are relaxed for Experiments 4-10 such that a failure is noted if a candidate is not found at least as good as the reference (AMR without noise suppression) at the 95% confidence interval ("equal or better than" criterion).

Simple Failures (excluding noise suppression in the downlink)	0	2	3	5	5	7
	1. NS5	**2. NS2**	**3. NS3**	**4. NS4**	**4. NS6**	**6. NS1**
Systematic Failures (excluding noise suppression in the downlink)	0	0	0	0	0	1
	1. NS2	**1. NS3**	**1. NS4**	**1. NS5**	**1. NS6**	**6. NS1**

Table 7.3: *Failures per candidate using Relaxed Performance Requirements*

It can be readily seen (Annex C) that all candidates have systematic failures in Experiment 10 for two special noise types: music noise and multiple interfering talkers. For additional information, the calculation of the number of failures (and the rankings) was carried out also for the case when music noise and multiple interfering talker noise (in Experiment 10) are excluded in the analysis, but where otherwise the Minimum Performance criteria of Table 5.1 are applied. This is justified by noting that it is not at all clear whether noise suppression functionality should attempt to suppress such background signals.

Simple Failures (excluding noise suppression in the downlink)	1	2	5	5	5	9
	1. NS5	**2. NS2**	**3. NS3**	**3. NS4**	**3. NS6**	**6. NS1**
Systematic Failures (excluding noise suppression in the downlink)	0	0	0	0	0	2
	1. NS2	**1. NS3**	**1. NS4**	**1. NS5**	**1. NS6**	**6. NS1**

Table 7.4: *Failures per candidate excluding Conditions with Music and Interfering Talkers*

7.3 Summary of Listening Test Results Covering Comparison of Candidates

This summary is presented in the form of tables of the FOMs defined in Section 6. It should be noted that SMG11 had stated that FOM#1 is the preferred Figure of Merit. This measures the ability of the solutions to suppress noise in terms of resulting speech quality compared to the non-suppressed speech. Having said this, it is clear that other Figures of Merit are significant

In particular note should be taken of FOM#7, derived from Experiment 3 which is used to detect unnatural effects in the noise-suppressed signal. In analysing the results according to FOM#10, it should be noted that Experiment 3 is designed to look for unnatural effects in the noise suppressed speech and turned out to be sensitive to distortions, which may cause the difference in the obtained FOM results compared to FOM#6. Experiment 3 has a large influence in FOM#10; hence the low and often negative values for this FOM.

None of the Figures of Merit listed below are intended to serve as a single selection criterion.

FOM#1	21.0962	17.0745	15.9298	15.5055	12.0572	12.0193
	NS6	**NS4**	**NS2**	**NS1**	**NS3**	**NS5**
FOM#3a	8.2708	6.8802	5.8854	5.7708	4.8906	4.7969
	NS6	**NS4**	**NS1**	**NS2**	**NS5**	**NS3**
FOM#3c	12.8253	10.1943	10.1590	9.6201	7.2603	7.1287
	NS6	**NS4**	**NS2**	**NS1**	**NS3**	**NS5**
FOM#4a	2.4167	1.9531	1.9167	1.6510	1.6302	1.5156
	NS6	**NS2**	**NS4**	**NS1**	**NS5**	**NS3**
FOM#4b	18.6795	15.1578	13.9767	13.8545	10.5416	10.3801
	NS6	**NS4**	**NS2**	**NS1**	**NS3**	**NS5**
FOM#6a	8.3969	7.9714	7.8701	6.6958	6.6818	5.9484
	NS6	**NS4**	**NS2**	**NS5**	**NS1**	**NS3**
FOM#6b	10.1798	8.0699	7.258	7.1044	5.824	4.2657
	NS6	**NS4**	**NS1**	**NS2**	**NS3**	**NS5**
FOM#6c	7.9647	6.2515	5.9248	4.5490	3.9159	2.85
	NS6	**NS1**	**NS4**	**NS2**	**NS3**	**NS5**
FOM#7a	6.5104	5.3229	4.8125	3.8021	3.4688	3.2917
	NS5	**NS2**	**NS4**	**NS3**	**NS6**	**NS1**
FOM#7b	4.2432	-2.3479	-9.1099	-20.4534	-31.8353	-42.9414
	NS5	**NS2**	**NS3**	**NS4**	**NS1**	**NS6**
FOM#10a	3.0387	1.7575	1.5281	1.4948	1.0221	0.127
	NS5	**NS4**	**NS2**	**NS1**	**NS3**	**NS6**
FOM#10b	0.641	-0.1528	-0.1573	-0.2158	-0.9801	-3.2787
	NS5	**NS1**	**NS4**	**NS2**	**NS3**	**NS6**
FOM#10c	-0.9665	-3.1681	-4.3471	-8.2638	-8.9911	-10.5122

	NS5	NS3	NS2	NS4	NS6	NS1

7.4 Graphical Representation of Results from all Formal Listening Tests

This section provides the results of all 14 sub-experiments from all labs in graphical form. For more detailed information see Annex C. Note these graphs have been imported directly from the Global Analysis spreadsheet, and therefore also contain data for noise suppression in tandem in the uplink and downlink (which formed part of the feasibility study).

The following abbreviations are used in conjunction with these graphs:

@x	Defined bit rate of AMR speech codec
AMR/NS	AMR with noise suppression active
DL	Downlink
T1	Single Connection, i.e. noise suppression present in the uplink only
T2	Tandem Connection, i.e. noise suppression present in the uplink and the downlink
UL	Uplink
w/DTX	with VAD/DTX active
w/tandem	Tandem connection (mobile to mobile) with noise suppression active in the uplink and downlink legs of the connection.

7.4.1 Experiment 2: Degradation in Clean Speech (pair comparison test)

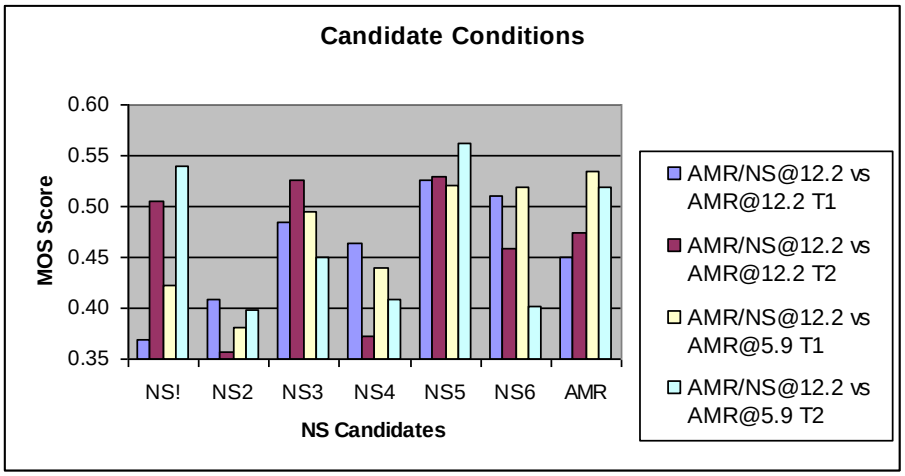

Figure 7.1: *Experiment 2 Results: English Language*

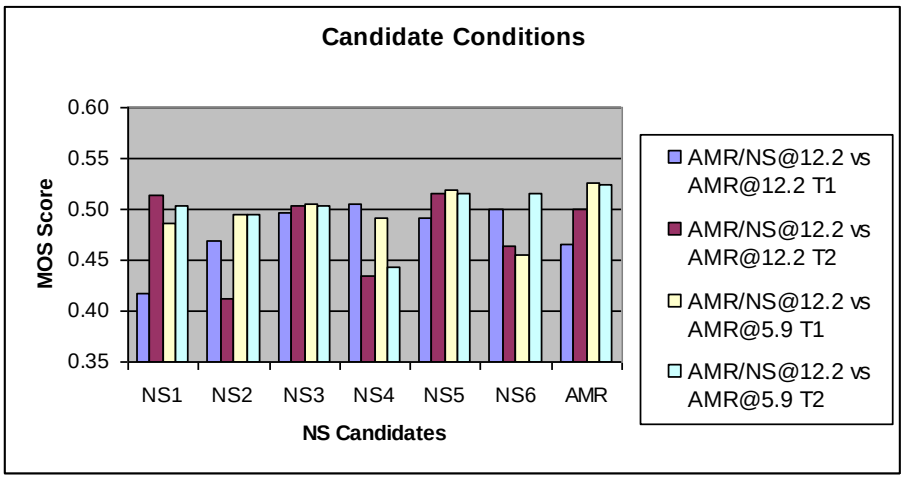

Figure 7.2: *Experiment 2 Results: French Language*

7.4.2 Experiment 3: Artifacts and Clipping in Background Noise

7.4.2.1 Car Noise

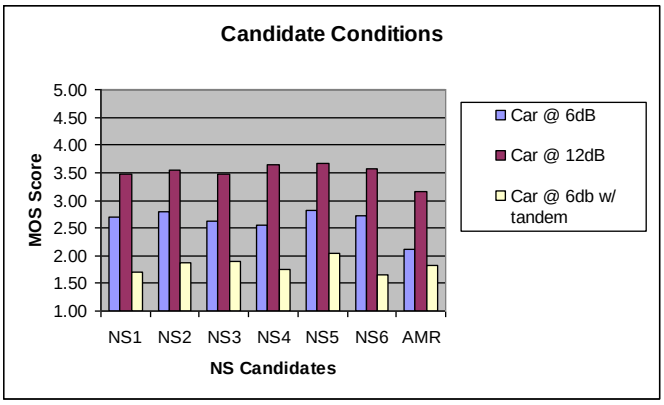

Figure 7.3: *Experiment 3 Results: Car Noise, English Language*

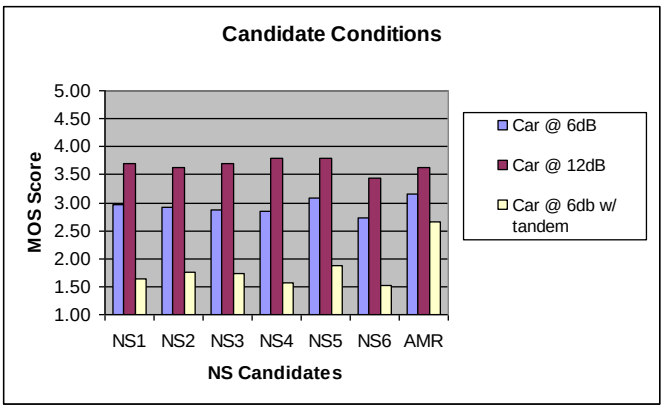

Figure 7.4: *Experiment 3 Results: Car Noise, Italian Language*

7.4.2.2 Street Noise

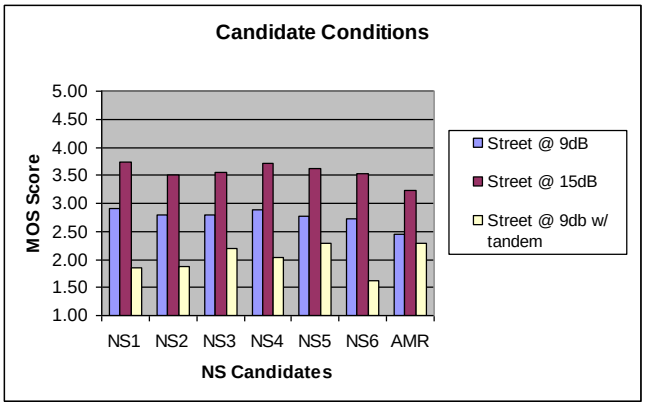

Figure 7.5: *Experiment 3 Results:Street Noise, English Language*

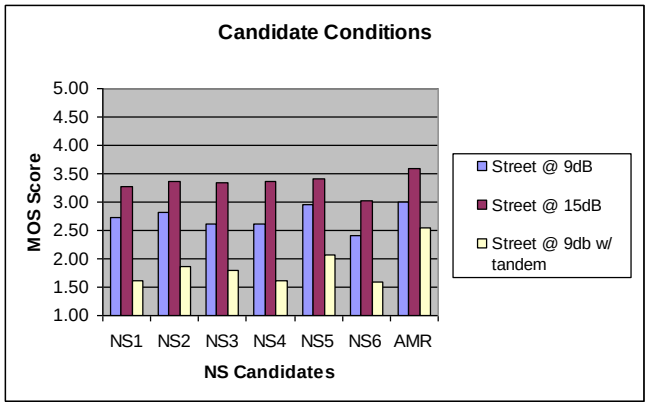

Figure 7.6: *Experiment 3 Results: Street Noise, Italian Language*

7.4.2.3 Babble Noise

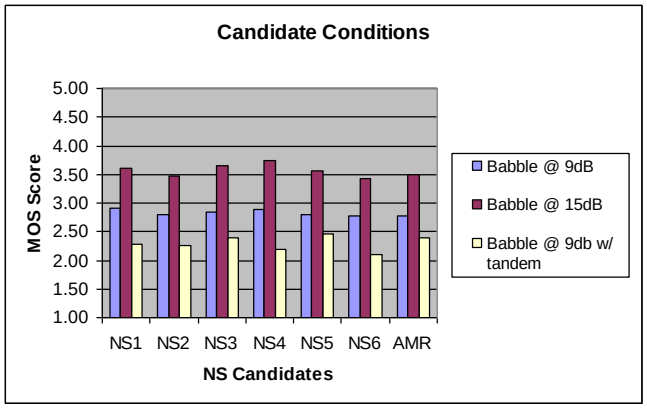

Figure 7.7: *Experiment 3 Results:Babble Noise, English Language*

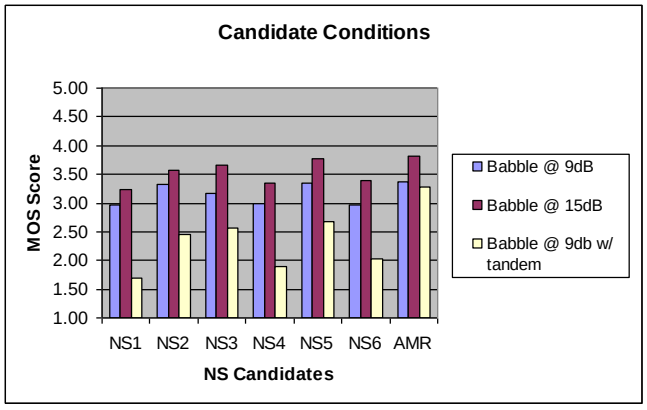

Figure 7.8: *Experiment 3 Results: Babble Noise, Italian Language*

7.4.3 Experiment 4: Performance in Background Noise (5.9kbps AMR Speech Codec)

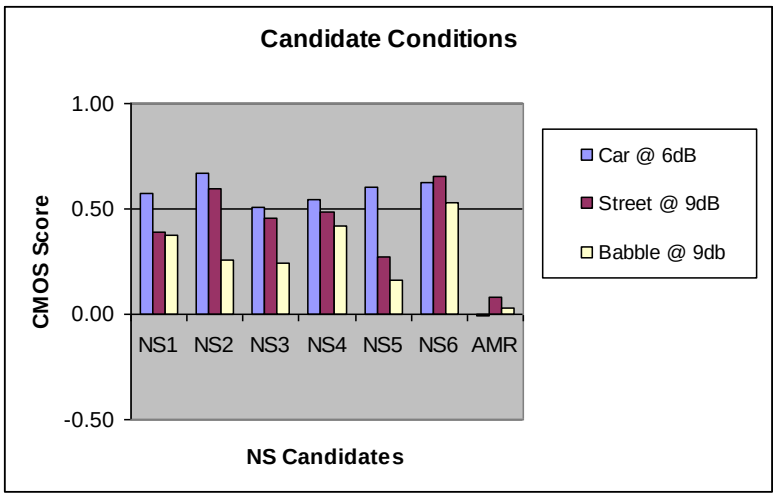

Figure 7.9: *Experiment 4 Results: Low SNR, English Language*

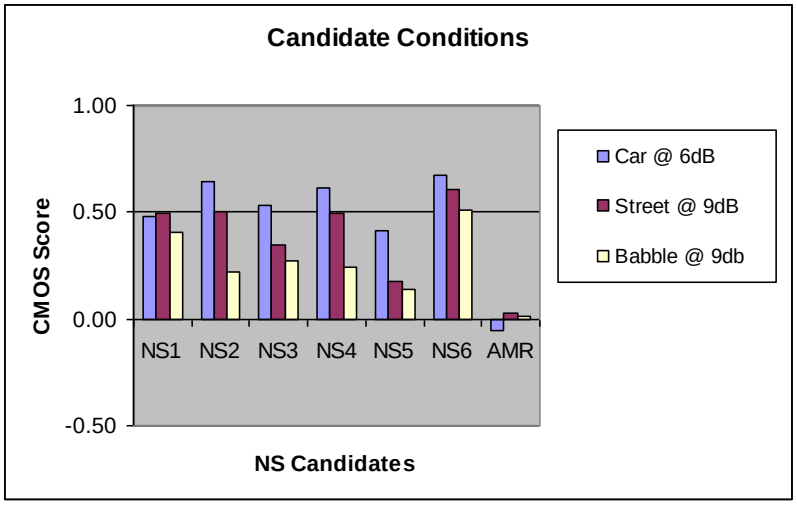

Figure 7.10: *Experiment 4 Results: Low SNR, Spanish Language*

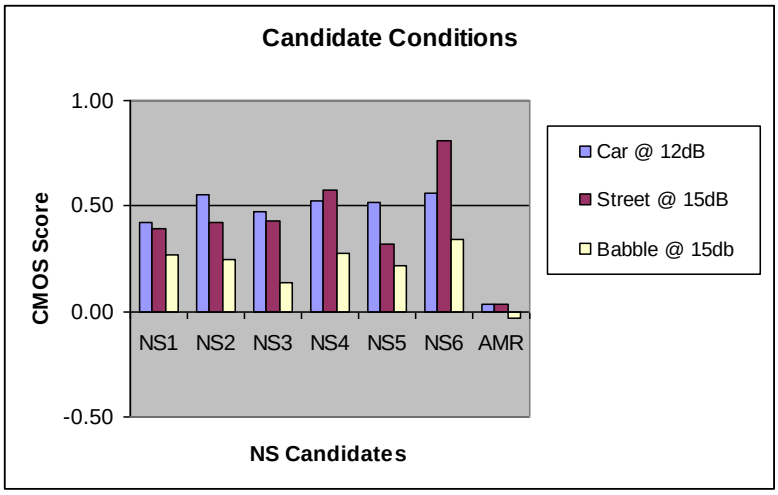

Figure 7.11: *Experiment 4 Results: High SNR, English Language*

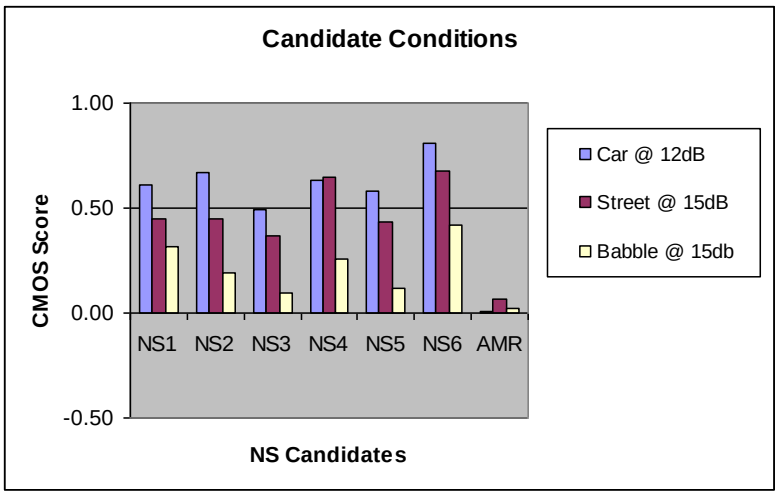

Figure 7.12: *Experiment 4 Results: High SNR, Spanish Language*

7.4.4 Experiment 5: Performance in Background Noise (12.2kbps AMR Speech Codec)

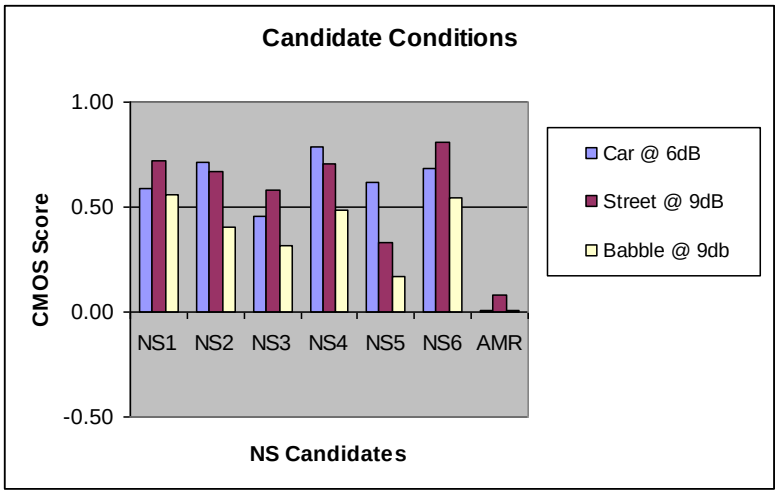

Figure 7.13: *Experiment 5 Results: Low SNR, English Language*

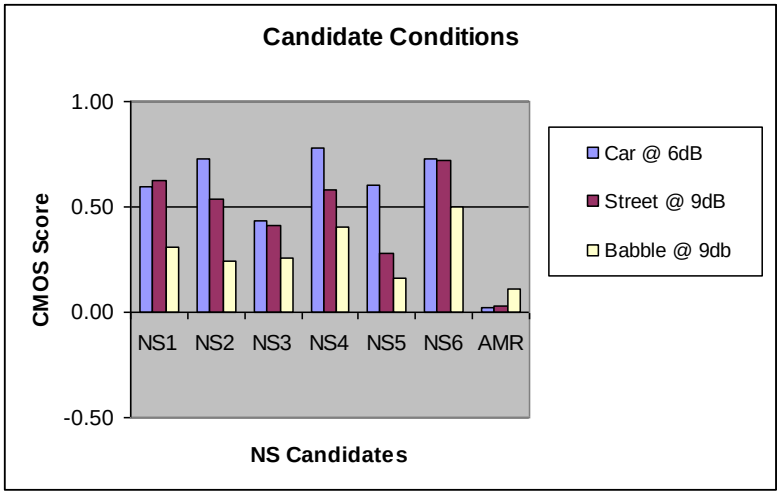

Figure 7.14: *Experiment 5 Results: Low SNR, Spanish Language*

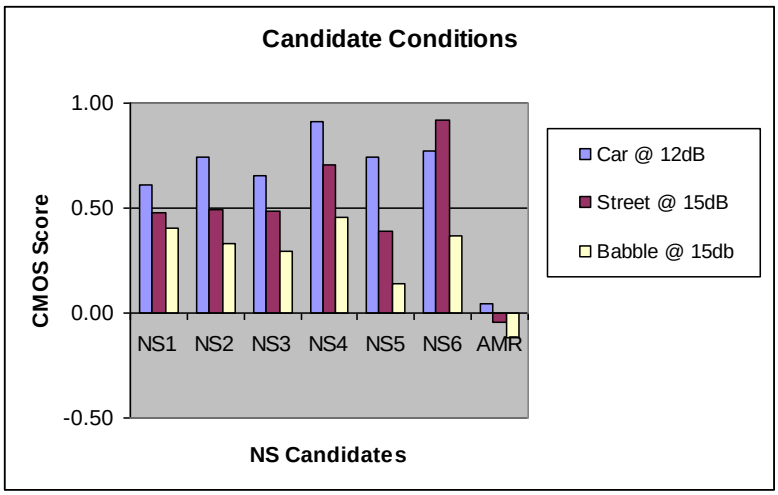

Figure 7.15: *Experiment 5 Results: High SNR, English Language*

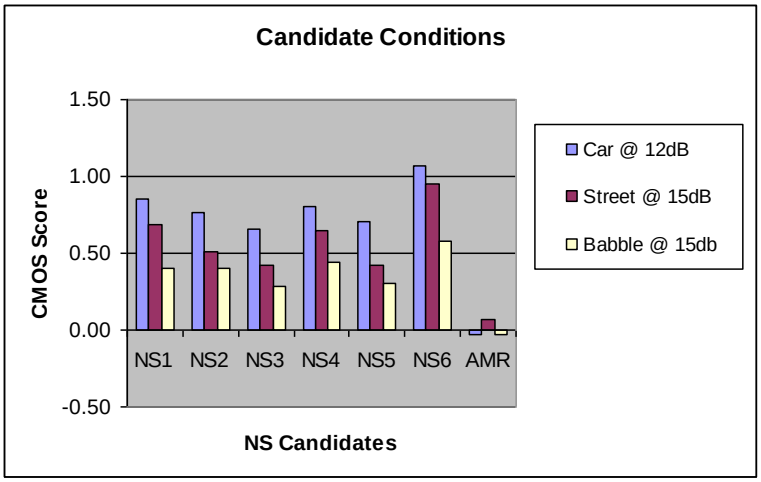

Figure 7.16: *Experiment 5 Results: High SNR, Spanish Language*

7.4.5 Experiment 6: Performance in Background Noise with Channel Errors (Car Noise with 6dB SNR)

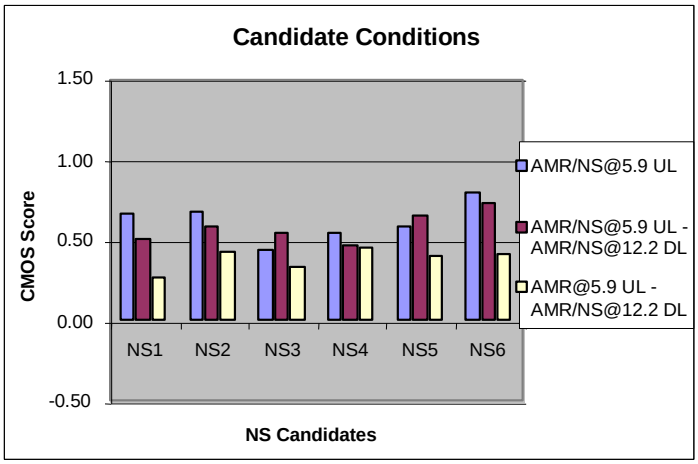

Figure 7.17: *Experiment 6 Results: English Language*

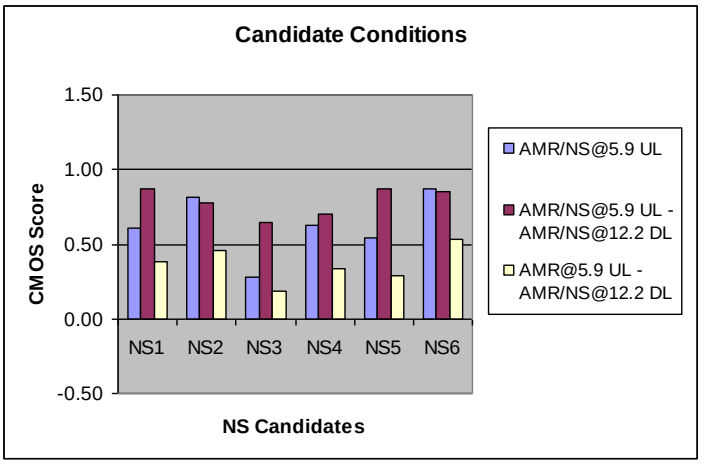

Figure 7.18: *Experiment 6 Results: Spanish Language*

7.4.6 Experiment 7: Performance in Background Noise with Channel Errors (Street Noise with 9dB SNR)

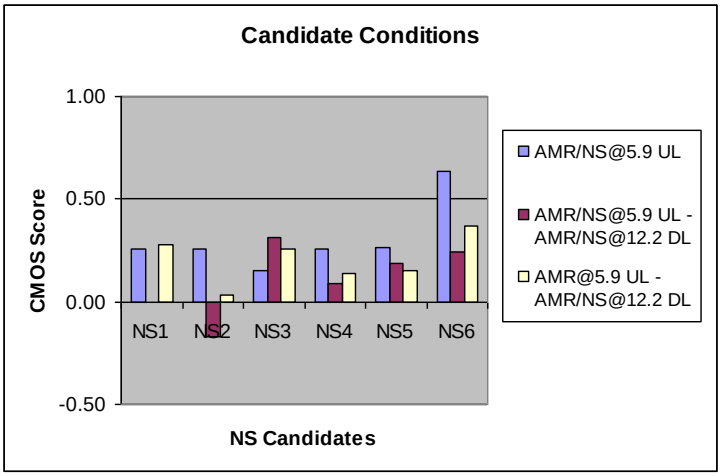

Figure 7.19: *Experiment 7 Results: English Language*

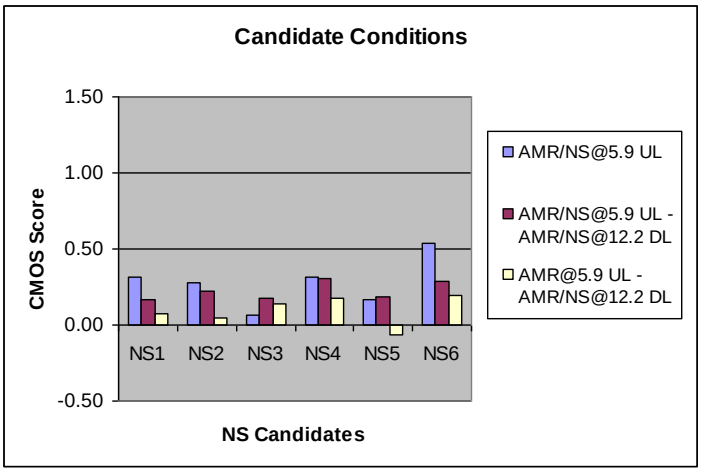

Figure 7.20: *Experiment 7 Results: Spanish Language*

7.4.7 Experiment 8: Performance in Car Noise with VAD/DTX active (VAD Option 1)

Note: The SNR for the car noise conditions in this experiment was set to 6dB.

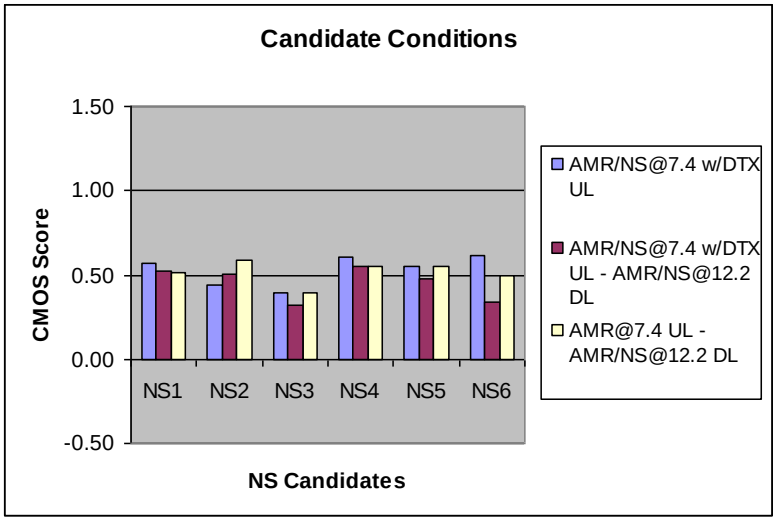

Figure 7.21: *Experiment 8 Results: English Language*

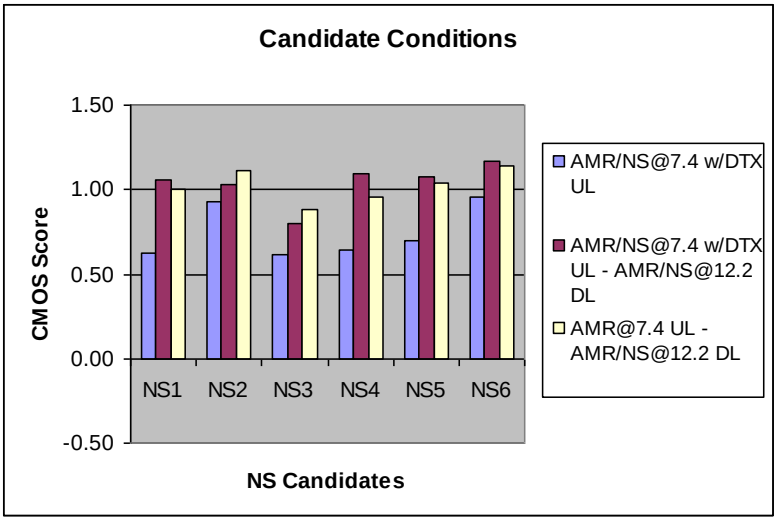

Figure 7.22: *Experiment 8 Results: Mandarin Language*

7.4.8 Experiment 9: Performance in Street Noise with VAD/DTX active (VAD Option 2)

Note: The SNR for the street noise conditions in this experiment was set to 9dB.

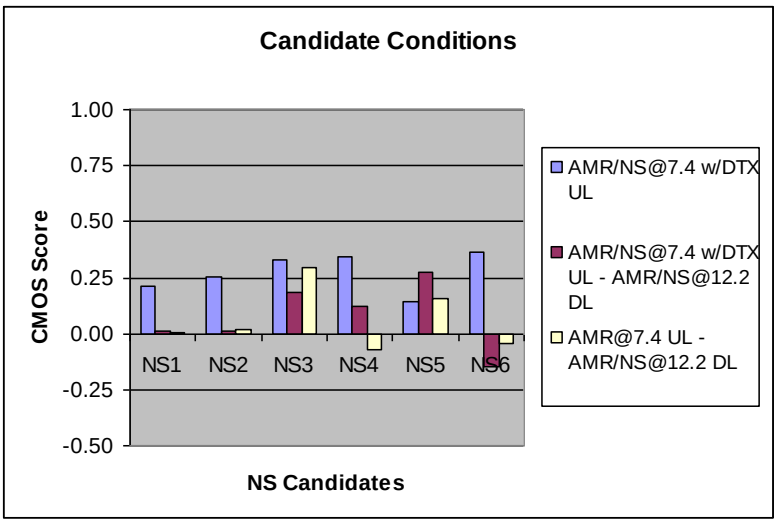

Figure 7.23: *Experiment 9 Results: English Language*

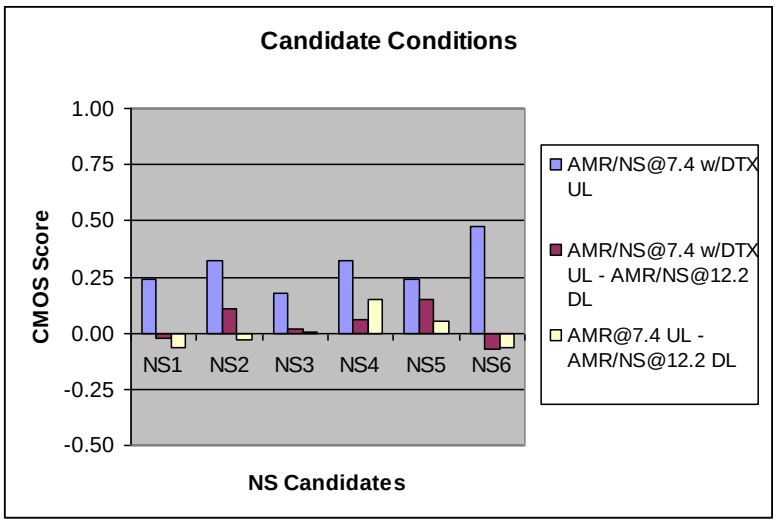

Figure 7.24: *Experiment 9 Results: Mandarin Language*

7.4.9 Experiment 10: Influence of Input Signal Level and Special Noise Types.

7.4.9.1. Influence of Input Level

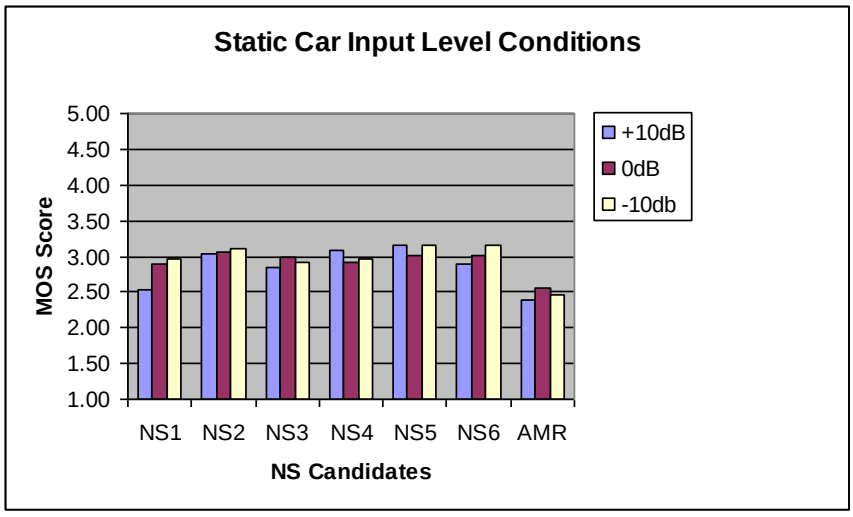

Figure 7.25: *Experiment 10 Results: Effect of Input Level, Car Noise, English Language*

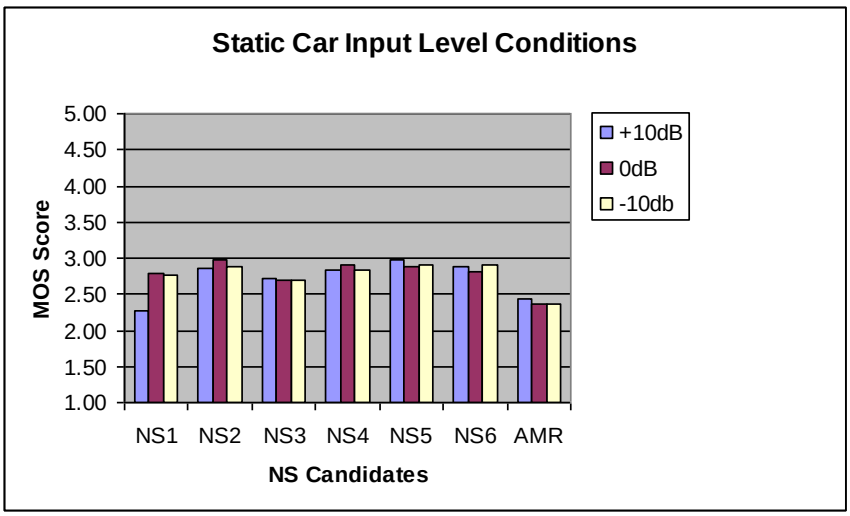

Figure 7.26: *Experiment 10 Results: Effect of Input Level, Car Noise, Japanese Language*

7.4.9.2 Performance with Special Noise Types

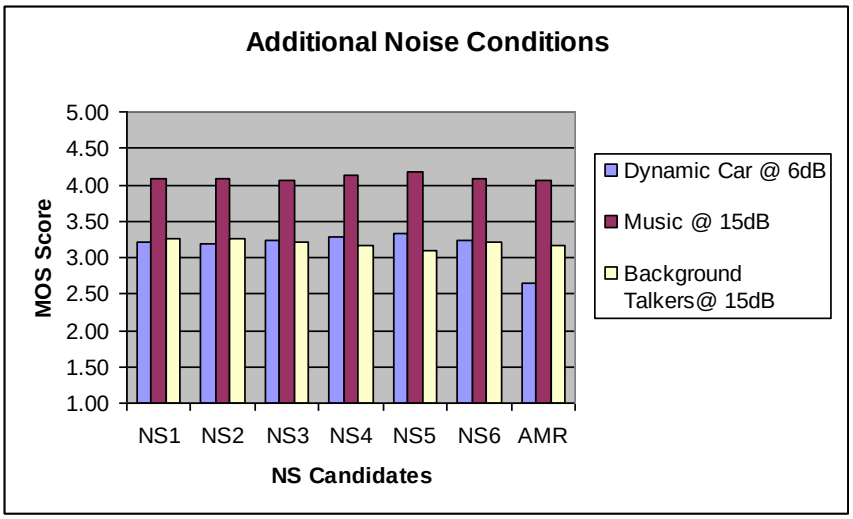

Figure 7.27: *Experiment 10 Results: Special Noises, English Language*

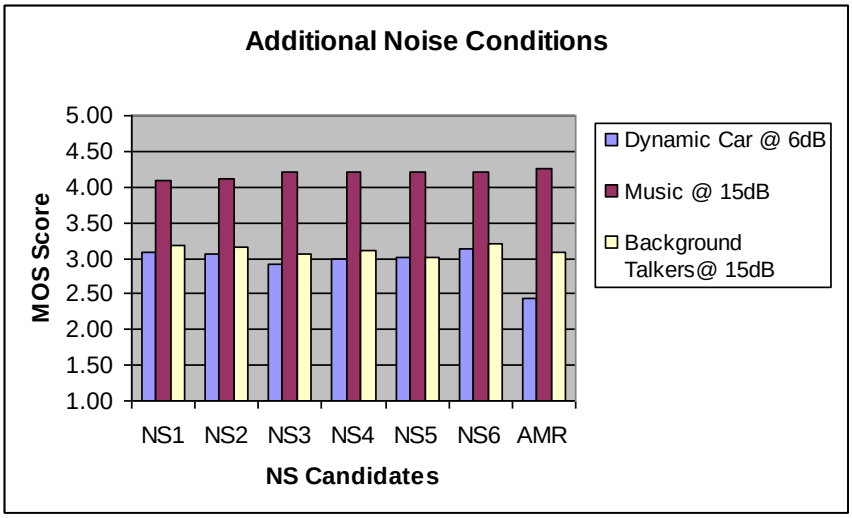

Figure 7.28: *Experiment 10 Results: Special Noises, Japanese Language*

8 Design Constraints

This section summarises the design constraints (limits on complexity, delay) and details the related values for all the candidates who took part in the Selection Phase.

Both the requirements (limits) and values for each candidate are provided in the Table 8.1.

In the context of this table, the following definitions are made. The DSP that runs the algorithm has been modelled through three parameters E, S and P. **E** stands for the Efficiency of the DSP. This corresponds to the ratio TMOPS/WMOPS of the implementation of the codec on the DSP. **S** stands for the Speed of the DSP: Maximum Number of Operations that the DSP can run in 1 second. This number is expressed in MOPS. **P** stands for the percentage of DSP processing power assigned to the codec. The processing delay of a task whose complexity is X can then be computed using the formula: D = X*20/ESP, the time unit being ms.

	NS1	NS2	NS3	NS4	NS5	NS6	Requirement
WMOPS	2,910	3,386	2,432	3,623	4,472	3,934	5,000
Dynamic RAM (words)	*770*	*2234*	*781*	*768*	*1529*	*1073*	*3039*
Static RAM (words)	262	718	168	577	850	239	1500
Data ROM (words)	312	863	302	731	877	537	1000
Program ROM (basic ETSI operations)	754	772	1018	907	884	581	2000
Delay (ms)	5,00	5,00	0,00	2,00	0,00	5,00	7,00
Delay-stand alone (ms)	5,00	5,00	1,50	10,75	0,00	5,00	
Implementation	*embedded*	*stand alone*	*embedded*	*embedded*	*stand alone*	*embedded*	

FOM(1)	5,72	8,49	5,38	8,33	10,34	6,48	15,80
FOM(2)@ESP25	7,33	7,71	1,95	4,90	3,58	8,15	11,00
FOM(2)@ESP50	6,16	6,35	0,97	3,45	1,79	6,57	9,00
FOM(2)@ESP100	5,58	5,68	0,49	2,72	0,89	5,79	8,00

FOM(1) = WMOPS + 2*sRAM + (2/5)* dROM + 2*pROM **sRam, dROM in kbytes, pROM in kbasic ETSI ops**

FOM(2) = delay(proc) + delay(algor) **delay(proc) = WMOPS * 20 /(E*S*P)); in ms**

Table 8.1: *Summary of Design Constraints Information.*

9 Impact on Voice Activity Factor VAF (with VAD/DTX active)

The Selection Phase Requirement concerning impact on VAD/DTX stated that the AMR speech codec with noise suppression activated should not significantly increase channel activity when used in conjunction with DTX.

Table 9.1 details the VAF increase for each candidate for each VAD option, as an average across all tested speech plus noise samples. In this table a positive value denotes an increase in VAF, whereas a negative value denotes a decrease in VAF.

Candidate	NS1	NS2	NS3	NS4	NS5	NS6
VAF increase for VAD Option 1 (%)	+1.77	+2.72	+0.11	-0.79	+0.20	+0.68
VAF increase for VAD Option 2 (%)	+0.09	+0.30	+0.00	-2.22	-0.42	+0.03

Table 9.1: *Summary of Impact on VAF*

10 Objective Performance Measurements

A tool was used to generate objective measures of performance (in terms of speech quality). This information is regarded as additional, and is in all cases secondary to the results obtained by subjective listening (as reported in Section 7). Two measures were undertaken on a subset of the material utilised in the listening tests. These were Noise Power Level Reduction (NPLR) and Signal to Noise Improvement (SNRI). Further details can be found in Annex E. The following tables provide the results of the analysis for each candidate, which details the NPLR results per noise type for each candidate.

	NS1	**NS2**	**NS3**	**NS4**	**NS5**	**NS6**
Car Noise	-8,43	-7,35	-7,53	-8,50	-8,40	-10,99
Street noise	-5,79	-2,23	-4,21	-5,75	-3,93	-5,37
Babble noise	-3,70	-0,47	-0,98	-2,42	-0,81	-0,78

Table 10.1: *NPLR Results Summary*

11 Feasibility Study: Downlink Noise Suppression for AMR

During the selection testing of the NS candidates, conditions including the noise suppression algorithm in the downlink path were tested. The aim was to assess the feasibility of putting the noise suppression algorithm in the network on the downlink path. Because the selection process was focused on the uplink, those conditions were not taken into account in the selection results. However, results are available and are noted here.

It was decided not to test the downlink path in isolation to avoid doubling the amount of testing required. Moreover, to perform a fair comparison, no different tuning of algorithm behaviour was allowed between the downlink and the uplink noise suppression algorithms.

The following table records the number of failures for each candidate in the conditions including noise suppression in the downlink (i.e. self-tandeming of the noise suppression algorithm). In total there were 26 conditions including noise suppression in the downlink.

Simple Failures (noise suppression in the downlink)	5	5	10	10	12	14
	1. NS3	**1. NS5**	**3. NS1**	**3. NS6**	**5. NS4**	**6. NS2**
Systematic Failures (noise suppression in the downlink)	0	0	3	3	4	5
	1. NS3	**1. NS5**	**3. NS1**	**3. NS4**	**1. NS6**	**6. NS2**

Table 11.1: *Failures per candidate for conditions including noise suppression in the downlink using the Minimum Performance Requirements*

Additionally results for the number of failures are presented in Table 11.2 where the requirements are relaxed for Experiments 6-9 such that a failure is noted if a candidate is not found at least as good as the reference at the 95% confidence interval ("equal or better than" criterion).

Simple Failures (noise suppression in the downlink)	3	3	4	6	7	7
	1. NS3	**1. NS5**	**3. NS1**	**3. NS6**	**5. NS2**	**6. NS4**
Systematic Failures (noise suppression in the downlink)	0	0	1	2	2	2
	1. NS3	**1. NS5**	**3. NS1**	**3. NS2**	**1. NS4**	**6. NS6**

Table 11.2: *Failures per candidate for conditions including noise suppression in the downlink using the Relaxed Performance Requirements*

The following table presents the FOMs defined for the cases with noise suppression in the downlink. FOM#5 is the summation of CMOS scores for all conditions in the CCR tests including noise suppression in the downlink. FOM#9a is the summation of all delta MOS scores for all conditions in the ACR tests including noise suppression in the downlink.

FOM#5	6.4739	6.4304	6.0918	5.7097	5.5772	5.5003
	NS5	**NS6**	**NS4**	**NS2**	**NS1**	**NS3**
FOM#9a	-1.5833	-2.3958	-2.8958	-3.8958	-4.1667	-4.4583
	NS5	**NS3**	**NS2**	**NS4**	**NS1**	**NS6**

Annex A: Key Selection Phase Documents

All the following documents can be found on the ETSI FTP site:

http://docbox.etsi.org/tech-org/smg/Document/smg11/SMG11_amr_ns/NS_Sel_Phase/

Design constraints:	AMR-NS_Design_Constraints1.0.doc
Selection Phase Deliverables:	Deliverables1-1.doc
Selection Rules:	NSSelRules1.1.doc
Processing functions	ProcFunc_v012.zip

Annex B: Selection Phase Test Plan

See associated file Test-plan.doc

Annex C: Global Analysis Spreadsheet

See associated files AMR-NS_CCR_v1.xls, AMR-NS_MOS_v1.xls

Annex D: Methodologies for Measuring Subjective SNR Improvement

D1: CCR Experiments

The purpose of experiments 4&5 is to evaluate the performances of the NS algorithms in background noise conditions with two different bit-rates (5.9 kbps and 12.2 kbps). For these experiments three types of noise have been selected: car noise, street noise and babble noise. For each type of noise two different nominal SNR levels have been set:

Noise type	SNR sub-exp. a [dB]	SNR sub-exp. b[dB]
Car	6	12
Street	9	15
Babble	9	15

For each sub-experiment and for each type of noise three (two for babble noise) ideal NS reference conditions will be processed:

Ideal SNR improvement
SNR sub-exp. +4 dB
SNR sub-exp. +7 dB
SNR sub-exp. +10 dB [1]

1 This condition will be available only for car and street noise

Each ideal NS will be compared during the sub-experiment with the speech+noise signals mixed at the nominal SNR levels. This lead to a total number of CCR reference results of 3 per sub-experiment (2 for the babble noise) corresponding to 3 (2 for the babble noise) SNR improvement levels. By connecting adjacent point by straight lines we will obtain a graph giving a correspondence between CCR notes and perceived SNR improvement (cf. figure D.1).

Finally the perceived SNR improvement for an AMR-NS candidate is obtained for each candidate using the CCR vs SNR graph as illustrated in figure D.1.

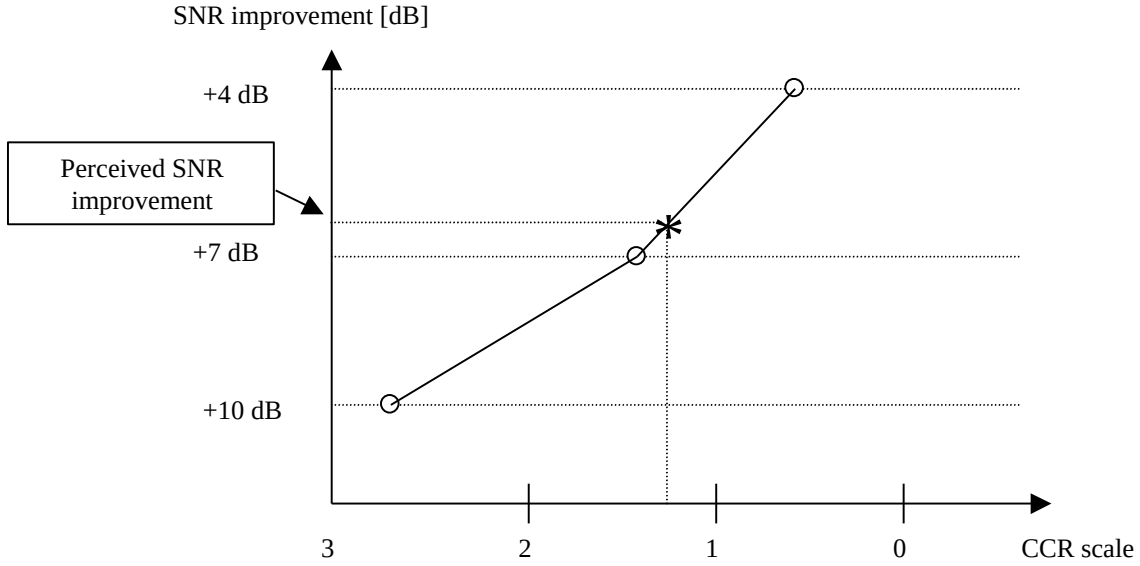

Figure D.1. Example of CCR versus SNR improvement graph
O: ideal NS score, *:AMR-NS candidate score.

Ranking for CCR experiments

The ranking of different algorithms is obtained by using a weighted sum of the perceived SNR improvement for each candidate according to:

$$\text{Score AMR-NSx} = \frac{1}{2}\left[0.6 \cdot (SNRimp\_4a + SNRimp\_5a) + 0.4 \cdot (SNRimp\_4b + SNRimp\_5b)\right]$$

where SNRimp_ny is the perceived SNR improvement for sub-experiment number ny. In this expression a higher weight is given to results obtained with a lower nominal SNR levels cause it is generally easier to discriminate the NS algorithms in the lower SNR.

D2: ACR Experiments

The methodology for evaluating the subjective SNR improvement for the ACR tests (Experiments 3 a, b, and c) is similar to the methodology used for the CCR tests. For each Experiment a, b, and c (car noise, street noise, and babble noise) the performance is evaluated for two different SNR levels, resulting in two sub-experiments per experiment:

Experiment	Noise type	SNR sub-exp 1 [dB]	SNR sub-exp 2 [dB]
3a	Car noise	6	12

3b	Street noise	9	15
3c	Babble noise	9	15

For each noise type in sub-experiment 1 (the lower SNR) the material will be processed with an ideal NS reference with attenuation of 4, 6, 8, 10, and 12 dB. For each noise type in sub-experiment 2 (the higher SNR) the material will be processed with an ideal NS reference of 4, 6, 8, 10 dB. (Note that some of the conditions in the sub-experiments will result in similar total SNR. However the speech sample randomization differs between the two sub-experiments). By connecting the ACR score for adjacent ideal NS reference attenuation points by straight lines, graphs giving correspondence between ACR scores and perceived SNR improvement is obtained (cf. Figure D.2) for each noise type in each sub-experiment. Similarly to the case for the CCR Experiments, the perceived SNR improvement for an AMR-NS candidate is obtained using the ACR vs SNR graph as illustrated in Figure D.2.

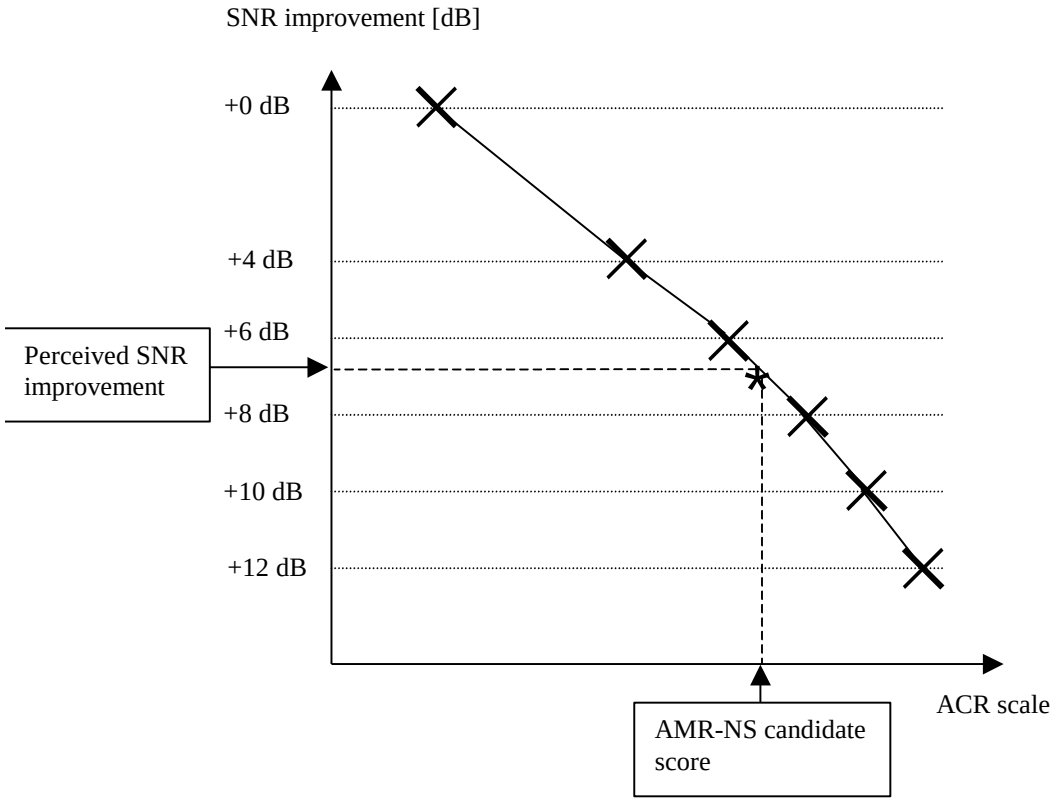

Figure D.2. Example of ACR versus SNR improvement graph
X: ideal NS score, *:AMR-NS candidate score.

Ranking for ACR experiments

The ranking of the different AMR-NS candidates for each noise type is obtained by averaging the subjective SNR improvement values for each of the two sub-experiments.

Annex E: Methodology for NS performance evaluation by Objective Means

This annex presents a two objective measures that were used in the AMR/NS selection phase for characterising the performance of the noise suppression (NS) candidate solutions.

NEW proposals for objective measures and TEST SIGNALS

Notations

The following notations are used in the formulation of the objective measures.

- The operator AMR(·) corresponds to applying the AMR speech encoder and decoder on the input.
- The operator NR(·) corresponds to applying the NS algorithm, and the AMR speech encoder and decoder on the input.
- The clean speech signals will be referred as $\mathbf{s_i}$ ***, i = 1 to I***.
- The noise signals will be referred as $\mathbf{n_j}$ ***, j = 1 to J***.
- The noisy speech test signals will be referred as $\mathbf{d_{ij} = \beta_{ij}(SNR)\, n_j + s_i}$***, i = 1 to I, j = 1 to J***, where d_{ij} is built by adding s_i and n_j with a pre-specified SNR as presented below.
- The processed signal will be referred as $\mathbf{y_{ij} = NR\,(d_{ij})}$, the operator NR(·) referring to the processing by the NS algorithm and the AMR speech codec.
- The reference signal in the calculations shall be either the noisy speech test signal d_{ij} itself or d_{ij} processed by the AMR speech codec without NS processing. The latter signal will be referred to as $\mathbf{c_{ij} = AMR\,(d_{ij})}$***, i = 1 to I, j = 1 to J***, where the operator AMR(·) refers to processing by the AMR speech codec with no NS. The relevant reference signal will be indicated in the formulation of each objective measure below.
- The notation *Log*(·) indicates the decimal logarithm.
- $\mathbf{\beta_{ij}(SNR)}$ is the scaling factor to be applied to the background noise signal $\mathbf{n_i}$ in order to have a ratio **SNR** (in dB) between the clean speech signal $\mathbf{s_i}$ and $\mathbf{n_j}$. The scaling of the input speech and noise signals is to be carried according to the following procedure:
  The clean speech material is scaled to a desired dBov level with the ITU-T recommendation P.56 speech voltmeter, one file at a time, each file including a sequence of one to four utterances from one speaker.
  A silence period of 2 s is inserted in the beginning of each of the resulting files to make up augmented clean speech files.
  Within each noise type and level, a noise sequence is selected for every speech utterance file, each with the same length as the corresponding speech files, and each noise sequence is stored in a separate file.
  Each of the noise sequences is scaled to a dBov level leading to the SNR condition corresponding to the $\mathbf{\beta_{ij}(SNR)}$ value in each of the test cases by applying the RMS level based scaling according to the P.56 recommendation.
- The determination of which frames contain active speech is to be carried out with reference to the ITU-T recommendation P.56 active speech level

measurement and is related to the classification of the frames into the presented speech power classes which is explained below.

Test material

The test material should manifest at least the following extent:

- Clean speech utterance sequences: 6 utterances from 4 speakers - 2 male and 2 female - totalling 24 utterances
- Noise sequences:
  car interior noise, 120 km/h, fairly constant power level
  street noise, slowly varying power level

Special care should be taken to ensure that the original samples fulfill the following requirements:

- the clean speech signals are of a relatively constant average (within sample, where 'sample' refers to a file containing one or more utterances) power level
- the noise signals are of a short-time stationary nature with no rapid changes in the power level and no speech-like components

Preferably, the test signals should cover the following background noise and SNR conditions:

- car noise at 3 dB, 6 dB, 9 dB, 12 dB and 15 dB
- street noise at 6 dB, 9 dB, 12 dB, 15 dB and 18 dB

A feasible subset of these conditions giving a practically useful indication of the achieved performance would be:

- car noise at 6 dB and 12 dB
- street noise at 9 dB and 15 dB

The samples should be digitally filtered before NS and speech coding processing by the MSIN filter to become representative of a real cellular system frequency response.

Note. In the application of the presented objective measures, there is no need to remove the 2 s initial convergence period referred to above after the processing from the test material. Namely, the classification of the frames being based on the clean speech signal and on comparisons to the active speech level, no frames from the initial convergence period will be involved in any of the measurements.

Proposal for objective measures for NS performance assessment

Assessment of SNR improvement level. The SNR improvement measure, ***SNRI***, measures the SNR improvement achieved by the NS algorithm. SNR improvement is calculated separately in three frame power gated factors of active speech signal, namely, high, medium and low power constituents of the signal. These categories are used to characterise the effect of the NS processing on speech, allowing to distinguish the effect on strong, medium and weak speech. In addition to calculating the SNR improvement separately on the three categories, they are used to form an aggregate measure.

The calculation is here presented for the high power speech class:

For each background noise condition j

For each speaker i

Construct a noisy input signal d_{ij} as follows:

$d_{ij}(n) = \beta_{ij}\ n_j(n) + s_i(n)$

where β_{ij} depends on the SNR condition according to the procedure described above

c_{ij} = AMR (d_{ij})

y_{ij} = NR (d_{ij})

$$\mathrm{SNRout\_h_{ij}} = \frac{\xi + \frac{1}{K_{sph}} \sum_{k=k_{sph,1}}^{k_{sph,K_{sph}}} \sum_{n=k\cdot 80}^{k\cdot 80+79} y_{ij}^2(n)}{\xi + \frac{1}{K_{nse}} \sum_{l=k_{nse,1}}^{k_{nse,K_{nse}}} \sum_{n=l\cdot 80}^{l\cdot 80+79} y_{ij}^2(n)} - 1$$

$$\mathrm{SNRin\_h_{ij}} = \frac{\xi + \frac{1}{K_{sph}} \sum_{k=k_{sph,1}}^{k_{sph,K_{sph}}} \sum_{n=k\cdot 80}^{k\cdot 80+79} c_{ij}^2(n)}{\xi + \frac{1}{K_{nse}} \sum_{l=k_{nse,1}}^{k_{nse,K_{nse}}} \sum_{n=l\cdot 80}^{l\cdot 80+79} c_{ij}^2(n)} - 1$$

$$\mathrm{SNRI\_h_{ij}} = \begin{cases} 0 & ; \quad \mathrm{SNRout\_h_{ij}} \le \xi \vee \mathrm{SNRin\_h_{ij}} \le \xi \\ 10 \cdot \left[\mathrm{Log}\left(\mathrm{SNRout\_h_{ij}}\right) - \mathrm{Log}\left(\mathrm{SNRin\_h_{ij}}\right) \right] & ; \quad \text{else} \end{cases} \qquad (1)$$

where k_{sph} and K_{sph} are the index and the total number of frames containing speech of a high power

k_{nse} and K_{nse} are the corresponding index and total number of noise only frames

$\xi > 0$ is a constant that should be set at 10^{-5}

$\mathrm{SNRI\_m_{ij}}$ correspondingly for medium power frames

$\mathrm{SNRI\_l_{ij}}$ correspondingly for low power frames

$$\mathrm{SNRI_{ij}} = \frac{1}{K_{sph} + K_{spm} + K_{spl}} \left(K_{sph} \cdot \mathrm{SNRI\_h_{ij}} + K_{spm} \mathrm{SNRI\_m_{ij}} + K_{spl} \mathrm{SNRI\_l_{ij}} \right)$$

(2)

$$\mathrm{SNRI_j} = \frac{1}{\mathrm{I}} \sum_{i=1}^{I} \mathrm{SNRI_{ij}} \qquad (3)$$

$$\mathrm{SNRI} = \frac{1}{\mathrm{J}} \sum_{j=1}^{J} \mathrm{SNRI_j} \qquad (4)$$

In addition, measures for the SNR improvement in the high, medium and low power speech classes (SNRI_h, SNRI_m, SNRI_l, respectively) shall be recorded based on the following formulae:

$$\mathrm{SNRI\_h} = \frac{1}{J} \sum_{j=1}^{J} \mathrm{SNRI\_h_j} = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{\mathrm{I}} \sum_{i=1}^{I} \mathrm{SNRI\_h_{ij}} \qquad (5)$$

$$\mathrm{SNRI\_m} = \frac{1}{J}\sum_{j=1}^{J}\mathrm{SNRI\_m_j} = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{\mathrm{I}}\sum_{i=1}^{I}\mathrm{SNRI\_m_{ij}} \tag{6}$$

$$\mathrm{SNRI\_l} = \frac{1}{J}\sum_{j=1}^{J}\mathrm{SNRI\_l_j} = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{\mathrm{I}}\sum_{i=1}^{I}\mathrm{SNRI\_l_{ij}} \tag{7}$$

To determine which frames belong to high, medium and low power classes of active speech and which present pauses in the speech activity (noise only), the active speech level (in dB) sp_lvl of the noise free speech $s_i(n)$ is first determined according to the ITU-T recommendation P.56. Thereafter, the frames are classified into the four classes as follows:

for all signal frames k

$$\mathrm{sp\_pow}(k) = 10\log\left[\max\left(\varepsilon, \frac{\sum_{n=k\cdot 80}^{k\cdot 80+79}(s_i(n))^2}{80}\right)\right] \tag{8}$$

if $\mathrm{sp\_pow}(k) \geq \mathrm{sp\_lvl} + \mathrm{th\_h}$

$$\{k_{sph,\mathrm{length}(k_{sph})+1}\} = \{k_{sph,\mathrm{length}(k_{sph})}, k\}$$

else if $\mathrm{sp\_pow}(k) \geq \mathrm{sp\_lvl} + \mathrm{th\_m}$

$$\{k_{spm,\mathrm{length}(k_{spm})+1}\} = \{k_{spm,\mathrm{length}(k_{spm})}, k\} \tag{9}$$

else if $\mathrm{sp\_pow}(k) \geq \mathrm{sp\_lvl} + \mathrm{th\_l}$

$$\{k_{spl,\mathrm{length}(k_{spl})+1}\} = \{k_{spl,\mathrm{length}(k_{spl})}, k\}$$

else if $\mathrm{sp\_lvl} + \mathrm{th\_nl} \leq \mathrm{sp\_pow}(k) < \mathrm{sp\_lvl} + \mathrm{th\_nh}$

$$\{k_{nse,\mathrm{length}(k_{nse})+1}\} = \{k_{nse,\mathrm{length}(k_{nse})}, k\}$$

where $\varepsilon > 0$ is a constant whose value shall be such that in the dB scale, it shall be below sp_lvl + th_nl; a value of 10^{-7} should be used if sp_lvl = -26 dBov and th_nl = -34 dB, as proposed below

th_h, th_m, th_l are pre-determined lower threshold power levels for classifying the speech frames to the high, medium, and low power classes, correspondingly.

We want to make the following notes on the formulation of the frame classification:

1. The lower bound for the power of the noise-only class of frames is motivated by a desire to restrict the analysis to noise frames that are among or close the speech activity, hence excluding long pauses from the analysis. This makes the analysis concentrate increasingly on the effects encountered during speech activity.
2. We realise that in poor SNR conditions, the noise power level may occur to be higher than the lower bound of some of the speech power classes. However, even in this case, the information of the effect on the low power portions of speech may be informative. Naturally, another way of formulating the measure might be to make the power thresholds dependent on the noise level. This would, however, restrict the comparability of the SNR improvement figures of the different classes over experiments with different background noise content.

3. The presented method of classifying the speech frames in the designated classes and, hence, determining values for the SNR improvement measures, is only applicable if all the used power level threshold values are higher than the corresponding power threshold level derived in the speech level measurement referred to above.

A preferable scaling for the clean speech material is a normalisation to the active speech level of –26 dBov. In such a case, the following values should be used for the power class thresholds:

th_h = -1 dB
th_m = -10 dB
th_l = -16 dB (10)
th_nh = -19 dB
th_nl = -34 dB

According to experimentation, the results of the analysis are not highly sensitive to the selection of the threshold values. However, the determination of the th_l and th_nh threshold values is somewhat critical to avoid confusion between low power speech and a weak background noise typically present in the clean speech samples.

Assessment of noise power level reduction. The noise power level reduction ***NPLR*** measure relates to the capability of the NS method to attenuate the background noise level.

The ***NPLR*** measure is calculated as follows:

For each background noise condition j
For each speaker i
Construct a noisy input signal d_{ij} as follows:
$d_{ij}(n) = \beta_{ij}\ n_j(n) + s_i(n)$
where β_{ij} depends on the SNR condition according to the procedure described above
c_{ij} = AMR (d_{ij})
y_{ij} = NR (d_{ij})

$$NPLR_{ij} = 10 \cdot \left\{ \mathrm{Log}\left[\xi + \frac{1}{K_{nse}} \sum_{k=k_{nse,1}}^{k_{nse,K_{nse}}} \sum_{n=k\cdot 80}^{k\cdot 80+79} y_{ij}^2(n) \right] - \mathrm{Log}\left[\xi + \frac{1}{K_{nse}} \sum_{l=k_{nse,1}}^{k_{nse,K_{nse}}} \sum_{n=l\cdot 80}^{l\cdot 80+79} c_{ij}^2(n) \right] \right\}, \quad (11)$$

where $\xi > 0$ is a constant, such as 10^{-5};
k_{nse} and K_{nse} are the corresponding index and total number of noise only frames

$$NPLR_j = \frac{1}{I} \sum_{i=1}^{I} NPLR_{ij} \quad (12)$$

$$NPLR = \frac{1}{J} \sum_{j=1}^{J} NPLR_j \quad (13)$$

Comparison of *SNRI* and *NPLR*. A comparison of the ***SNRI*** and ***NPLR*** measures can be used to acquire an indication of possible speech distortion produced by the tested NS method. If the ***NPLR*** parameter assumes clearly higher values than ***SNRI***, it can be expected that the NS candidate causes distortion to speech. This relation, however, always needs to be verified through a comparison between the objective measures and corresponding subjective test results.

Comments on the AMR/NS selection test material

We have expressed above the premise that the street noise test material used in conjunction with the presented objective quality measures should be of a slowly varying power level. As a candidate proponent having gone through the processing of the source speech material with our AMR/NS candidate solution, we now have some experience on the noise material used for the AMR/NS selection tests. Our impression of the street noise material is not quite consistent with the requirement stated above. Namely, the street noise samples appear to contain, to some extent, background speech, horns and similar components whose frame power varies in a rate whose range coincides that of speech. Hence, the results to be obtained for the street noise conditions have to be interpreted with special care.

On the scope of usage of objective measures for NS evaluation

The objective measures presented in this document are intended for characterising some relevant aspects of the performance of NS algorithms. Prior to the selection phase testing, it was noted that they might help in the comparison of AMR/NS candidates that are found equal by other means. However, we want to emphasise that the figures obtained with the proposed measures were decided to be used as auxiliary information only. The subjective test results were acknowledged as the principal data for ranking the AMR/NS candidates in the selection process.

Annex F: Methodology for Measuring Impact on Voice Activity Factor (VAF)

This contribution presents a proposal for the measurement of the VAF (Voice Activity Factor) references. This method is also suggested for the measurements of the candidate's VAF. Nortel Networks will conduct them and provide a report for the Noise Suppresser Selection Phase (i.e. 15th Nov. 1999). Nortel Networks will also provide the means to conduct the same measurements for a cross checking.

References :

[1] Noise Suppression for the AMR codec, Service Description, Stage 1.

[2] Tdoc SMG11 288/99, Test Plan Specification for the AMR NS Selection Phase v1.7.

2.0 Voice activity Factor Measurement

2.1 General

The Voice Activity Factor is defined as the ratio of the number of frames declared as speech (SPEECH) by the AMR Voice Activity Detector (VAD) over the total number of frames during a given time.

The parameter of interest for an operator regarding the radio usage efficiency is the mean Radio Channel Activity Factor (RAF) in a cell. This RAF corresponds to the ratio of the number of transmitted bursts to the number of timeslots available during a given time. The RAF is somehow linked to the VAF (depending on the Traffic channel FR or HR, the number of SID_FIRST frames and the number of SID_UPDATE frames). For the sake of simplicity, we limit the measurement to the VAF. But, the method described and the C code also enable the computation of the RAF if needed.

2.2 VAF requirements

The requirements for the NS candidate regarding the VAF are the following [1]:

"The AMR speech codec with noise suppression activated should not significantly increase channel activity when used in conjunction with DTX.

Channel activity increase will be measured thanks to the Voice Activity factor (VAF), defined as follows.

Let x be the VAF measured by the AMR VAD as an averaged value on all clean speech signals

Let y be the VAF measured by the AMR VAD without AMR NS active as an averaged value on all clean speech + noise signals (where the applicable clean speech signal is the speech signal used in the measure of x).

Let w be the VAF measured by the AMR VAD with AMR NS active as an averaged value on all clean speech +noise signals (where the applicable clean speech signal is the speech signal used in the measure of x). w is required to be less than the maximum of y and x. Any case where w is greater than y should be further investigated.

For real world signals, w is required not to be significantly greater than y. Any case where w is greater than y should be further investigated.

These requirements shall apply to all standardised AMR VADs. (w,x,y) are determined using all VADs, and the requirements are checked relatively to each AMR VAD independently."

As a consequence, values X and Y are independent of the NS candidate. They are considered as reference values. W can be computed by the candidate using the same procedure and compared to X and Y. The calculation of Y (resp. X) is described in section 2.3 (resp. 2.4).

X, Y and W values should be compatible in the sense that the original speech material shall be the same for all of them. There should also be no speech material used twice.

We propose to measure the VAF by counting SPEECH frames in the output file of the AMR encoder. Therefore the preprocessed noisy speech material provided by ARCON and COMSAT will be used for Y value. The corresponding preprocessed files without added noise will be used for X. The process will exclude propagation error conditions, synthesis and tandeming (i.e. the only process will be the AMR encoding stage with or without the NS activated). All downlink conditions are excluded from the process since the VAF requirements are only applicable to uplink. Therefore, when the original processing in the test plan includes both up and downlink, only the uplink processing is done for the VAF.

The first two seconds used for convergence are included for the processing but the computation of the VAF should ignore those two seconds.

It has to be noted that no real worlds signals were included in the test plan.

Based on this we can make the following remarks :

- No speech files will be concatenated.
- No specific weighting will be applied to files with respect to experiment since the requirement doesn't separate noise types. Anyhow, the program outputs the VAF for each type of noise and this might be subject to analysis if needed.
- No specific weighting will be applied to files with respect to the AMR Mode. The mode used for the processing before the VAF measurement is the same as the one used in the processing test plan.
- The same speech files may be used twice in two different conditions if the encoding mode is different for each condition.

2.3 Measurement for Noisy speech (Y)

It has been agreed that the VAF measurement will be done using all the speech material used for the selection testing. Since X and Y are measured using the same material for comparison reasons, only noisy speech are retained. The following table lists the experiments potentially used for the computation.

Exp. No.	Title	No. of Sub-Exp.
3	Artefacts, Clipping & Distortion Effects in Background Noise Conditions	3
4 & 5	Performances in Background Noise Conditions	4
6 & 7	Influence of Propagation Error Conditions	2
8 & 9	Influence of Voice Activity Detection and Discontinuous Transmission	2
10	Influence of the Input Signal +Noise Level and Performances with Special Noise Types	1

After having excluded double usage of speech material with the same encoding mode, we end up with the following list for the preprocessed files provided by ARCON:

Exp. No.	Conditions retained	Excluded conditions
3a	19, 25	31
3b	19, 25	31
3c	19, 25	31
4a	15, 21, 27	
4b	15, 21, 27	
5a	15, 21, 27	
5b	15, 21, 27	
6a	7, 13(uplink only)	19
7a	7, 13(uplink only)	19
8a	7, 13(uplink only)	19
9a	7, 13(uplink only)	19
10a	13, 19, 25, 31, 37, 43	

2.4 Measurement for Clean speech (x)

For clean speech signals, the same procedure will be used. Therefore the corresponding material should be preprocessed without adding noise samples

2.5 Measurement for Noisy speech with the candidate NS algorithm activated (w)

As for the previous case, the same procedure will be used. The C code given as an attached file can be used by candidates to perform their own measurements using their NS candidate. The advantage would be that the output values will correspond to the reference values and that we will be able to do an "apple to apple" comparison during the selection.

3.0 Summary of the C code for VAF measurements

The attached Zip file contains the following source files :

- **Main.c**
  loops on speech files to do the process and the VAF measurements.

  Writes the report file.

- **Process.c**

  Process the speech file (AMR encoder with DTX on and NS optionally) and measures the number of SPEECH frames and the number of total frames in the output bitstream.

- **Desc.c**

  Contains the preprocessed file description to enable simple looping on speech files.

- **Main.h**

  Contains function prototypes and user-defined values.

The make file is not provided. The code was successfully tested using a PC/Windows 95 environment with Visual C++.

The user must change the following values according to its needs :

in file main.h:

```
initial dir of ARCON Files
ARCON_BASE              "d:/hlaba"
initial dir of COMSAT Files
COMSAT_BASE             "e:/hlabc"
Command line specific to the candidate
for VAD option 1
COMMAND_LINE_VAD1       "encoder [-ns_on] -dtx"
for VAD option 2
COMMAND_LINE_VAD2       "encoder2 [-ns_on] -dtx"
Report filename
REPORT_FILENAME         "VAF_Report.txt"
Candidate Acronym
```

ARCON_CANDIDATE_ACRONYM "xx"

COMSAT_CANDIDATE_ACRONYM "xx"

in file desc.c

Group of files descriptions can be changed to match the file structure

The program processes each file according to the descriptor array file. The processing follows the provided command line. The resulting file of the process is analyzed in order to count various frame types in the bit-stream ignoring the first 100 frames (2 seconds). The values of interest are returned to the main function that performs total and means calculation and writes the report. An example of report (with dummy data) is also attached to this proposal.

Annex G (informative): Change history

SA4# 25bis	Tdoc SA 4	Spec	CR	Cat	PH	Vers	New Vers	Subject
	S4-030264					8.0.0	8.0.1	Addition of VAF C-Code tool (Courtesy Nortel Networks)