3GPP TS 06.77 V8.1.1 (2001-04)

Technical Specification

3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Minimum Performance Requirements for Noise Suppresser Application to the AMR Speech Encoder

(Release 1999)

Keywords
GSM, codec, performance

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 Specification 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

The present document specifies recommended minimum performance requirements for noise suppression algorithms intended for application in conjunction with the AMR speech encoder. This specification is for guidance purposes. Noise Suppression is intended to enhance the speech signal corrupted by acoustic noise at the input to the AMR speech encoder.

The use of this recommended minimum performance requirements specification is not mandatory except for those solutions intended to be endorsed by SMG11.

It is the intention of SMG11 to perform analysis and validation of any AMR noise suppression solution which is voluntarily brought to the attention of SMG11 in the future, using the requirements set out in this specification to facilitate such an analysis. In order for SMG11 to endorse such a solution, SMG11 must confirm that all the recommended minimum performance requirements are met.

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] CCITT Recommendations I.130 (1988): "General modelling methods - Method for the characterisation of telecommunications services supported by an ISDN and network capabilities of an ISDNî".

[2] 3GPP TR 01.04 (ETR 350): "Digital cellular telecommunications system (Phase 2+); Abbreviations and acronyms".

[3] 3GPP TS 06.71 "Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR); speech processing functions; General description" Release 98.

3 Definitions and abbreviations

GSM 01.04 (ETR 350) [2] provides a list of abbreviations and acronyms used in GSM specifications. For the purposes of the present document, the following definitions and abbreviations also apply:

3.1 Definitions

None

3.2 Abbreviations

AMR	Adaptive Multi-Rate
AMR/NS	Combination of the AMR speech codec and the Noise Suppression function
NS	Noise Suppression

4 Description of Noise Suppression applied to AMR

Noise Suppression for the AMR codec is a feature designed to enhance speech quality in a range of environments where there is significant (acoustic) background noise. The noise suppression function is a pre-processing module that is used to improve the signal to noise ratio of a speech signal prior to voice coding. In so doing it may use functions and/or data from the AMR speech encoding function. This specification defines recommended minimum performance requirements for such a function when it is implemented in the mobile station (operating on the uplink speech signal).

The AMR Speech decoder should not be altered by the Noise Suppression function.

It shall be possible to disable the operation of the noise suppression algorithm using signalling when commanded by the network.

4.1 Applicability of Noise Suppression to Basic Services.

This feature shall be applicable (as an option) to all speech calls where the narrowband AMR codec is utilised. Provision of the feature in AMR-capable mobile stations is a manufacturer dependent option. The network shall be able to enable or disable this noise suppression function both at call set-up and in call. Signalling between network and mobile to allow this control has been provided.

5 Requirements to be assessed by Objective Means

5.1 Bit Exactness of the Speech Encoder

The Noise Suppression shall be implemented as a separate pre-processing module prior to the speech encoding. The functionality and all internal states, tables and variables of the speech encoder shall remain unaltered by the Noise Suppression function.

The Noise Suppression should be implemented as a stand-alone pre-processing module operating on the 160 samples input speech buffer to the speech encoder according to Figure 1.

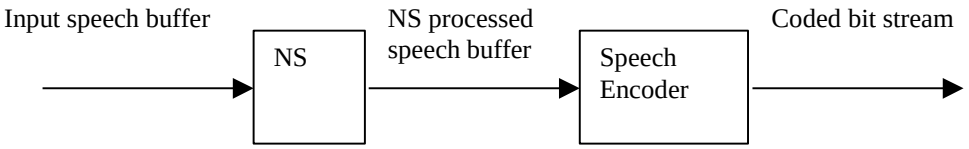

Figure 1: Noise Suppression implementation

Alternatively, for implementation in conjunction with the bit-exact fixed point C reference code [GSM 06.73] the NS module may operate on the pre-processed input speech buffer "old_speech[L_TOTA L]" in the structure "cod_amrState" in the AMR C code [GSM 06.73] after the pre-processing module (sample down-scaling and input high pass filtering) of the speech encoder. The bit-integrity of the speech encoder for this implementation shall be verified according to Figure 2 where the signals at Test Points 1 and 2 shall be identical for any input signal and the Reference Encoder is the part of [GSM 06.73] after the pre-processing module. Note: implementation in conjunction with the AMR floating point C code is for further study.

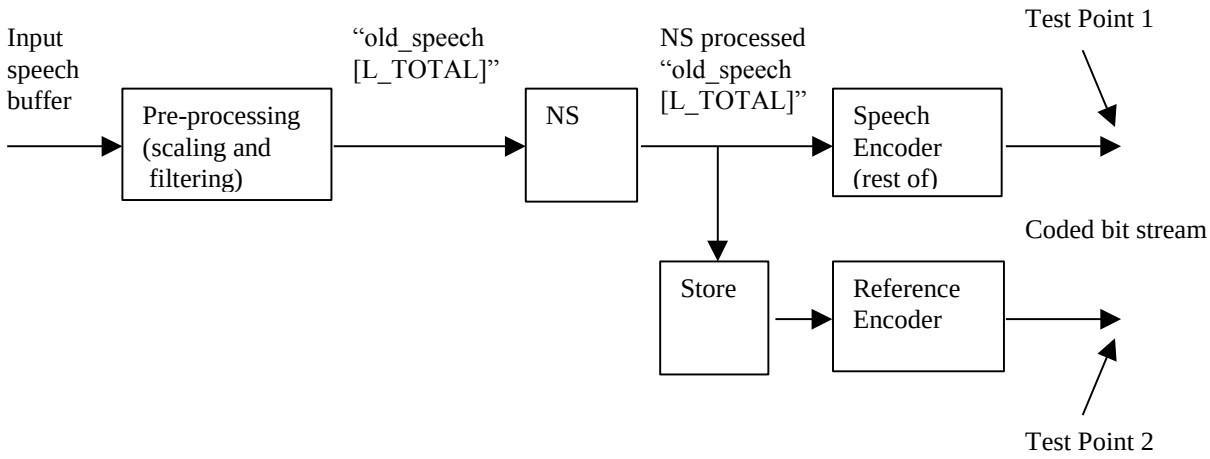

Figure 2: Verification of AMR speech encoder bit-exactness for embedded NS implementations

5.2 Bit Exactness of the Speech Decoder

The AMR speech decoder shall remain unaltered by the Noise Suppression function.

5.3 Impact on Speech Path Delay

The one way algorithmic delay due to the activation of AMR noise suppression shall be no more than 5ms in excess of the delay inserted by the AMR speech codec. In the handsfree case, this delay is part of the 39ms delay specified in GSM 03.50.

The total additional delay (comprising of algorithmic and processing delays) shall not exceed 10ms. The processing delay is calculated using the following formula with E*S*P set to 50.

$$delay(proc) = WMOPS * 20/(E * S * P)$$

where WMOPS = complexity in weighted operations per second evaluated through the theoretical worst case. (Direct means of measurement of total delay is for further study.).

5.4 Impact on Channel Activity

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 not significantly more than the maximum of y and x. Any case where w is greater than y should be further investigated.

These requirements shall apply to both standardized AMR VADs. (w,x,y) are determined using one or both VADs, and, if both are used, the requirements are checked relatively to each AMR VAD independently.

The definition of upper limits on VAF increase and attendant confidence intervals are for further study.

6 Requirements to be assessed by subjective tests

6.1 Impact on Speech Quality

The following performance requirements are stated under the assumption that the noise suppresser is tested as an integral part of the AMR speech codec with the speech codec operating at the rates defined within the test plan (Annex C) .The performance requirements must be met for all these stated speech codec rates.

6.1.1 Initial Convergence Time

The initial convergence time shall be a maximum of T seconds with T equal to 2s. The definition of this time interval shall be understood strictly in accordance with its means of use in subjective listening experiments. Its use shall be defined by a process whereby the first T seconds of each sample processed through the AMR speech codec with and without noise suppression active, is deleted before presentation to listeners. It is assumed that this process does not reduce intelligibility, or introduce clipping or similar effects into the resultant speech plus noise material.

6.1.2 No Degradation in Clean Speech

The noise suppression function must not have a statistically significant distorting effect on clean speech, in comparison with the performance of the AMR codec without noise suppression applied. This requirement also applies when VAD/DTX is active.

The requirement is checked with the use of a paired comparison test where the requirement is met if AMR/NS is preferred or equal to AMR within the 95 % confidence interval.

6.1.3 No degradation of Speech and no Undesirable Effects in Residual Noise in Conditions with Background Noise (*residual noise = background noise after AMR/NS)*

The noise suppression function must not introduce any degradation of speech and no undesirable effects in the residual noise, when there is (acoustic) background noise in the speech signal. This requirement also applies when VAD/DTX is active.

The requirement is checked with the use of a modified ACR test with specific instructions where the requirement is met if AMR/NS is better than or equal to AMR within the 95 % confidence interval in all conditions.

6.1.4 Quality Impact compared to AMR

The AMR speech codec with noise suppression activated must produce an output in noisy speech which is preferred amongst test listeners with statistical significance, compared to the case where noise suppression is not used. This requirement also applies when VAD/DTX is active.

The requirement is checked with the use of a CCR test where the requirement is met if AMR/NS is preferred to AMR within the 95 % confidence interval in at least 4 of the 6 conditions tested. Preference or equality within the 95 % confidence interval is required for the remaining conditions.

Additionally, it is required that the subjective SNR improvement as measured by the methodology [Annex B] (where the measure is conducted on the associated CCR tests [Annex C] meets the following requirements:

(a) In at least 2 of the 6 conditions tested the SNR improvement shall not be less than 6dB within the 95% confidence interval.

(b) In at least 2 of the remaining 4 conditions the SNR improvement shall not be less than 4dB within the 95% confidence interval.

7 Performance Objectives assessed by Objective Measures

7.1 Impact on Active Speech Level

The AMR speech codec with noise suppression activated must not significantly alter the active speech level.

The requirement is checked with the use of a P.56 speech level meter (the use of which remains for further study). Let x be the averaged level of the clean speech material for one experiment and let y be the averaged level of the processed material with AMR NS activated for the same experiment. The requirement is met if the absolute difference between x and y is less than 2 dB for all experiments. *The processed material should not be normalised to the nominal speech level before the listening tests.*

Note that this requirement does not preclude the use of active level control.

7.2 Objective Speech Quality Measures

The objective measures of noise power level reduction (NPLR) and signal-to-noise ratio improvement (SNRI) defined in Annex 1 are to be used to characterise the performance of the AMR/NS solution. Objectives are defined for these measures in the following table. These measures will be used to provide additional information only and are not to be considered to be requirements.

C source code is attached to this specification which shall be used to undertake these measurements.

Objective quality measure/test condition	**Performance objective**
NPLR *Assessment*: To be evaluated using a predefined set of material (as used in the AMR/NS Selection Phase) comprising speech mixed with stationary car noise in the SNR conditions of 6 dB and 15 dB, following otherwise the guidelines set forth in [Annex 1].	–7 dB or lower
SNRI *Assessment:* To be evaluated using a predefined set of material (as used in the AMR/NS Selection Phase) comprising speech mixed with stationary car noise in the SNR conditions of 6 dB and 15 dB, following otherwise the guidelines set forth in [Annex 1].	6 dB or higher

8 Interaction with supplementary services

8.1 General

This clause defines requirements regarding the interactions between GSM supplementary services and the Noise Suppression Feature.

The application of Noise Suppression shall not interfere with the provision or invocation of any supplementary services.

8.2 Explicit Call Transfer (ECT)

No adverse interaction. If the new party is a mobile station with support for the Noise Suppression feature, the noise suppression feature shall be invoked.

8.3 Call wait/Call hold.

No interaction.

8.4 Multiparty

No interaction.

8.5 Service Announcements

No interaction.

9 Interaction with Alternate and Followed by services

There shall be no impact on data transmission due the Noise Suppression Feature

10 Interaction with other speech services

There is no requirement for Noise Suppression in ASCI services.

11 Interaction with DTMF and other signalling tones

DTMF and other signalling tones transmission performance during the application of Noise Suppression shall be no worse than the case where Noise Suppression is turned off.

12 Interaction with Lawful Intercept

In the case where lawful intercept is required in a call where Noise Suppression is activated, the Noise Suppression shall not cause any degradation in the speech quality received by the A and B parties.

13 Interaction with TFO

No interaction.

Annex A (informative): Method for generating Objective Performance Measures

This annex presents an objective methodology for characterising the performance of noise suppression (NS) methods. Two objective measures are specified to be used for characterising NS solutions complying with the AMR/NS specification.

A.1 Notations

The following notations are used in this document:

- 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 are referred to as s_i, i = 1 to I.
- The noise signals are referred to as n_j, j = 1 to J.
- The noisy speech test signals are referred to as $d_{ij} = \beta_{ij}(SNR)\, n_j + s_i$, 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 are referred to as $y_{ij} = NR(d_{ij})$.
- 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 $c_{ij} = AMR(d_{ij})$, i = 1 to I, j = 1 to J. The relevant reference signal will be indicated in the formulation of each objective measure below.
- The notation *Log*(·) indicates the decimal logarithm.
- $\beta_{ij}(SNR)$ is the scaling factor to be applied to the background noise signal n_i in order to have a ratio **SNR** (in dB) between the clean speech signal s_i and n_j. The scaling of the input speech and noise signals is to be carried according to the following procedure:
    1. 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.
    2. A silence period of 2 s is inserted in the beginning of each of the resulting files to make up augmented clean speech files.
    3. 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.
    4. Each of the noise sequences is scaled to a dBov level leading to the SNR condition corresponding to the $\beta_{ij}(SNR)$ 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.

A.2 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

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 15 dB
- street noise at 9 dB and 18 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.

A.3 Objective measures for characterization of NS algorithm effect

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 groups of frames that represent power gated constituents of active speech signal. Hence, the *SNRI* measure is calculated separately in frames of high, medium and low power. 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. A frame length of 80 samples is used since it has been found the most efficient to describe changes in the signal caused by NS processing.

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:

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

where β_j depends on the SNR condition according to the procedure described above

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

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

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

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

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

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:

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

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

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

It is, in addition, informative to record separately the noise type specific SNR improvement measures, namely, SNRI_hj, SNRI_lj, SNRI_mj and SNRlj for each j.

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 . Let us first define four number sequences: $\{k_{sph}\}$, $\{k_{spm}\}$, $\{k_{spl}\}$, $\{k_{nse}\}$. All four sequences are initialized to an empty sequence:

- $$\begin{aligned}\{k_{sph}\}_0 &= \varnothing\\ \{k_{spm}\}_0 &= \varnothing\\ \{k_{spl}\}_0 &= \varnothing\\ \{k_{nse}\}_0 &= \varnothing\end{aligned} \quad (8)$$

Then, the frame power is calculated in each signal frame *k*:

$$\mathrm{sp\_pow}(k) = 10\log\left[\max\left(\varepsilon, \frac{\sum_{n=k\cdot 80}^{k\cdot 80+79}(s_i(n))^2}{80}\right)\right] \quad (9)$$

We shall then classify each frame according to the frame power as follows:

$$\begin{aligned}&\text{if } \mathrm{sp\_pow}(k) \geq \mathrm{sp\_lvl} + \mathrm{th\_h}\\ &\quad \{k_{sph}\}_{\mathrm{length}(k_{sph})+1} = \left\{\{k_{sph}\}_{\mathrm{length}(k_{sph})}, k\right\}\\ &\text{else if } \mathrm{sp\_pow}(k) \geq \mathrm{sp\_lvl} + \mathrm{th\_m}\\ &\quad \{k_{spm}\}_{\mathrm{length}(k_{spm})+1} = \left\{\{k_{spm}\}_{\mathrm{length}(k_{spm})}, k\right\}\\ &\text{else if } \mathrm{sp\_pow}(k) \geq \mathrm{sp\_lvl} + \mathrm{th\_l}\\ &\quad \{k_{spl}\}_{\mathrm{length}(k_{spl})+1} = \left\{\{k_{spl}\}_{\mathrm{length}(k_{spl})}, k\right\}\\ &\text{else if } \mathrm{sp\_lvl} + \mathrm{th\_nl} \leq \mathrm{sp\_pow}(k) < \mathrm{sp\_lvl} + \mathrm{th\_nh}\\ &\quad \{k_{nse}\}_{\mathrm{length}(k_{nse})+1} = \left\{\{k_{nse}\}_{\mathrm{length}(k_{nse})}, k\right\}\end{aligned} \quad (10)$$

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. In the following, these threshold values are called *power class threshold values*

$\mathrm{length}(k)$ is a function returning the length of the number sequence $\{k\}$

The following notes on the formulation of the frame classification are made:

- 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.

- 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. 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.

The scaling for the clean speech material should be determined optimally so that the dynamics of the 16 bit arithmetic system is efficiently used but no waveform clipping is produced. Typically, a normalisation to the active speech level of –26 dBov is preferable. In such a case, the following values should be used for the power class thresholds:

$$\begin{aligned} th\_h &= -1\ dB \\ th\_m &= -10\ dB \\ th\_l &= -16\ dB \\ th\_nh &= -19\ dB \\ th\_nl &= -34\ dB \end{aligned} \qquad (11)$$

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:

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

where β_j depends on the SNR condition according to the procedure described above

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

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

$$NPLR_{ij} = 10 \cdot \left\{ \mathrm{Log}\left[\xi + \frac{1}{K_{nse}} \sum_{m=1}^{K_{nse}} \sum_{n=k_{nse,m}\cdot 80}^{k_{nse,m}\cdot 80+79} y_{ij}^2(n) \right] - \mathrm{Log}\left[\xi + \frac{1}{K_{nse}} \sum_{l=1}^{K_{nse}} \sum_{p=k_{nse,l}\cdot 80}^{k_{nse,l}\cdot 80+79} c_{ij}^2(p) \right] \right\}, \qquad (12)$$

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

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

$$NPLR_j = \frac{1}{\mathrm{I}} \sum_{i=1}^{I} NPLR_{ij} \quad (13)$$

$$NPLR = \frac{1}{\mathrm{J}} \sum_{j=1}^{J} NPLR_j \quad (14)$$

Furthermore, it is informative to record separately the noise type specific NPLR measures, or $NPLR_j$, for each j.

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 absolute values than ***SNRI***, it can be expected that the NS candidate causes distortion to speech. This relation, however, should always be verified through a comparison with subjective test results.

Annex B (normative): Methodology for Measuring Subjective SNR Improvement for CCR Experiments

The purpose of experiment 3 is to evaluate the performances of the NS algorithm 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 [dB]
Car	6, 15
Street	9, 18
Babble	9,18

For each sub-experiment and for each type of noise three ideal NS reference conditions will be processed. The exception is that for the higher SNRs (15dB for car noise and 18 dB for street, babble noise) only 2 ideal noise reference conditions will be tested (+3, +6dB):

Ideal SNR improvement
SNR sub-exp. +3 dB
SNR sub-exp. +6 dB
SNR sub-exp. +9dB

Each ideal NS will be compared during the sub-experiment with the speech+noise signals mixed at the nominal SNR levels. This leads to a total number of CCR reference results of 5 per sub-experiment corresponding to 3 (2 for the higher SNRs) SNR improvement levels. By connecting adjacent point by straight lines we will obtain a graph giving a correspondence between CCR scores and perceived SNR improvement (cf. figure B.1).

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

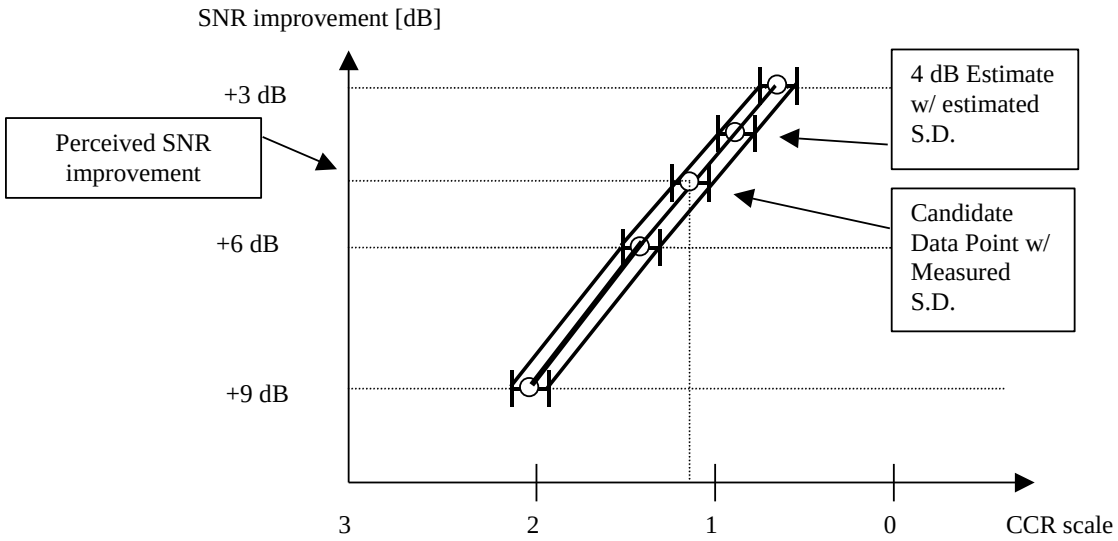

Figure B.1. Example of CCR versus SNR improvement graph
O: ideal NS score, *:AMR-NS candidate score.

Annex C (normative): Test Plan for Checking Conformance to Requirements

Document History:

Issue 0.1	16 Mar 00	First Issue, derived from the AMR/NS Selection Test Plan version 2.2 (Tdoc. SMG11/S4 356/99 R3)
Issue 0.2	28 Mar 00	D. Pascal (France Télécom R&D): Text of Experiment 1+ Annex A.1; Various editorial modifications
Issue 0.3	31 Mar 00	S. Aftelak (Motorola): Insertion of section on CCR tests + Annex A.3
Issue 0.4	23 May 00	S. Aftelak (Motorola): Mainly additions/changes to CCR tests
Issue 0.5	19 June 00	D. Pascal (France Télécom R&D): Statistical Analysis for Experiment1 (PC test)
Issue 0.6	05 Sept 00	Experimental Table for CCR Experiment 4
Issue 0.7	19 Oct 00	A. Eriksson (Ericsson): Addition of ACR tests
Issue 0.8	25.Oct.00	[SQ/Osaka]: • Note added to Section 8.11 on instructions • Added placeholder for statistical analysis section for Exp.4 • Added modified ACR instructions in Annex A
Issue 0.9	27 Oct 00	A Eriksson (Ericsson) • Editorial errors in Exp 2 and Exp 3 corrected • Experiment 4 changed to ACR test
Issue 0.10	16 Jan 01	A Eriksson (Ericsson) • Experiment 2 extended with odd level and DTX conditions • Experiment 4 reverted to previous CCR experiment
Issue 0.11	18 Jan 01	D. Pascal (France Télécom R&D) : Various editorial corrections and details, Statistical analysis for CCR experiments 3 and 4 is not correct and should be modified.
Issue 0.12	22 Jan 01	S Aftelak (Motorola): Editorial changes and changes to reflect fact that experimenter is responsible for providing (and reporting) processing tables and randomizations used.
Issue 2.0.0	22 Jan 01	Agreed at S4#15 Plenary meeting.

C1. Introduction

This document contains the complete set of subjective test experiments for the testing of the speech performance of Noise Suppression solutions for application to AMR. The purpose of the tests is to check for compliance to the recommended minimum performance requirements [1].

The AMR-NS Selection Tests are split into 4 main Experiments and 7 Sub-Experiments listed in the following table.

Exp. No.	Title	No. of Sub-Exp.
1	Degradation in Clean Speech (PC)	1
2	No degradation of Speech and no Undesirable Effects in Residual Noise in Conditions with Background Noise (ACR)	3
3	Performances in Background Noise Conditions (Mod-CCR)	2
4	Influence of Input Level, Voice Activity Detection and Discontinuous Transmission (Mod-CCR)	1
	Total Number of Sub-Experiments:	7

C2 Document Structure

The main body of the document starts at section 4, and is arranged as follows:

Section 4:	References, Conventions, and Contacts	References to specification documents, lists of abbreviations, and contact names for the different areas of the document
Section 5:	Roles and Responsibilities	Identification of roles and allocation of Responsibilities.
Section 6:	Information Relevant to all Experiments	Information relevant to all experiments.
Sections 7-10:	Test Plans	Individual test plans. Information already covered in section 6 is not repeated in the individual plans. Note that the processing tables for the experiments are collated in Annex B, and the randomizations (where required) in Annex C
Annex A:	Instructions to Subjects and Data Collection	For the Modified CCR, Pair Comparison, Modified ACR.
Annex B:	Processing Tables	Processing Tables for all experiments. These map which speech samples are to be processed through which conditions.
Annex C:	Presentation Orders	Randomized presentation orders for experiments.

C3. References, Conventions, and Contacts

[1]	GSM 06.77	Minimum Performance Requirements for Noise Suppresser Application to the AMR Speech Encoder (latest version)
[2]	TBD	Processing Function for the GSM AMR Noise Suppresser Selection Tests (Proposal - re-use selection phase document)
[3]	ITU-T Com 12	Handbook on Telephonometry
[4]	ITU-T Rec. P.800	Methods for subjective determination of transmission quality
[4]	GSM 06.71	Adaptive Multi-Rate Speech Codec; General Description
[5]	GSM 06.73	Adaptive Multi-Rate Speech Codec; ANSI C-Code
[6]	GSM 06.75	Performance Characterization of the GSM Adaptive Multi-Rate Speech Codec
[7]	GSM 06.90	Adaptive Multi-Rate Speech Codec; Transcoding Functions
[8]	GSM 06.91	Adaptive Multi-Rate Speech Codec; Error Concealment of Lost Frames
[9]	GSM 06.92	Adaptive Multi-Rate Speech Codec; Source Controlled Rate Adaptation
[10]	GSM 06.94	Adaptive Multi-Rate Speech Codec; Voice Activity Detector

C4a Key Acronyms

ACR	Absolute Category Rating
AMR	Adaptive Multi-Rate Speech Codec for the GSM System
AMR-NS	Noise Suppresser for the AMR Speech Codec
BER	Bit Error Rate
C/I	Carrier to Interference Ratio
DCR	Degradation Category Rating
DECi	Dynamic Error Condition #i for Dynamic C/I conditions
ECx	Error Condition for static C/I conditions with C/I = x dB
EFR	GSM Enhanced Full Rate speech codec
EP	Error Pattern
FR	GSM Full Rate channel or existing GSM Full Rate speech codec
HR	GSM Half Rate channel or existing GSM Half Rate speech codec
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
S/N	Signal to Noise Ratio

C4b Contact Names

The following persons should be contacted for questions related to the test plan.

Section	Contact Person/Email	Organization	Address	Telephone/Fax
Overall				
Experiments 1	Dominique Pascal/ dominique.pascal@rd.francetelecom.fr	France Télécom R&D	2 Av. Pierre Marzin Technopole Anticipa 22307 Lannion Cedex France	Tel : + 33 2 96 05 15 78 Fax : + 33 2 96 05 13 16
Experiments 2	Anders Eriksson/ anders.eriksson@era-t.ericsson.se	Ericsson		
Experiments 3	Steve Aftelak/ Stephen.Aftelak@motorola.com	Motorola		
Experiment 4	Steve Steve Aftelak/ Stephen.Aftelak@motorola.com	Motorola		

C5 Roles and Responsibilities

It is the sole responsibility of the proponent of a noise suppression solution to ensure that the testing is conducted properly according to this test plan. It is strongly recommended that third party subjective testing laboratories be instructed to perform the tests according to this plan.

It is additionally required that the test material is processed in accordance with the processing functions document [2].

Each experiment should be conducted in at least 2 languages. The proponent of the noise suppression proponent is at liberty to choose the languages to be used, but it is recommended that a reasonable range of languages be incorporated, across the full set of experiments.

C6 Information relevant to all Experiments

C6.1 General Technical Notes

Any and all deviations from the specifications contained in this document and the Processing Functions document [2] must be documented and submitted to SMG11/S4 along with the experimental results.

C6.2 Codec Adaptation and Error Conditions

The philosophy of the AMR system is that it is capable of dynamically altering the ratio of speech and channel coding to maximize speech performance as channel conditions change. Each of the combinations of speech and channel coding rates is known as a mode.

However, for the purpose of the AMR Noise Suppresser tests, only fixed mode operation will be considered.

C6.3 Speech Material

All AMR-NS Experiments are subjective listening experiments using pre-recorded speech passed through the candidate algorithms and simulated impairment conditions prior to use in the experiments. Three types of speech sample are used in these experiments:

- Single sentence samples, 4 seconds in length
- Short samples; sentence pairs, 8 seconds in length.
- Long samples; sentence quadruplets, 16 seconds in length.

The experiment investigating the equivalence of the candidate Noise Suppresser algorithms to the AMR algorithm without noise suppression in a quiet environment (PC experiment 1) will use the single sentence stimuli. The experiments investigating the possible introduction of artifacts and clipping by the candidate Noise Suppresser algorithms (ACR experiments 2a, 2b & 2c) will use the long 16-second samples. Experiment 2 includes conditions investigating level dependency, VAD and DTX. All other experiments will use the short 8-second samples.

For all original speech samples a 2s header will be added to accommodate the Initial Convergence Time of the Noise Suppresser algorithms. For all experiments this header should be removed at the end of the processing prior to being used in subjective listening tests.

Information for constructing these sentences is provided in the remainder of this subsection.

Unless stated otherwise in the individual plans, each source speech file will contain unique speech material (i.e. none of the sentences used in any given sample should be used in any other sample for the same, or any other talker within any sub experiment).

Pre-recorded source speech material may possibly be purchased as described in Section 6.3.1. Preferably, the test house should provide its own source speech material. The guidelines contained in Section 6.3.2 should be followed.

To avoid noise contrast effects, any silence gaps and/or pauses added to the speech files to pad them out into the specified formats for the source speech samples described in sections 6.3.3, 6.3.4 and 6.3.5, should not be pure digital silence. Padding out should be done by adding the ambient noise present during the recording of the speech material between the sentences.

The information in sections 6.3.3, 6.3.4 and 6.3.5 should be used in the preparation of the material that the talkers will utter, as well as how the recorded material should be constructed.

C6.3.1 Availability of Pre-recorded Speech Material

A "Multi-lingual Speech Database for telephonometry 1994", on 4 CD-ROM disks, was available from NTT-AT, No.7 Hakuei Buildg, 2-4-15 Naka-machi, Musashino-shi, 180 Japan (phone: +81 422 37 0823, fax: +81 422 60 4806).

In this database, the speech samples consist of pairs of short sentences with a total length of 8-10 seconds. Each sentence lasts approximately 2 to 3 seconds. Four male and four female native speakers are assigned to each of the 21 languages and 96 speech samples are available for each language. The sampling rate is 16 kHz. Active speech level (as defined in ITU-T Rec. P.56) of every speech sample is adjusted to -26dBovl.

Each CD consists of two different areas: audio and data. Speech samples in the audio area are digitized by 44.1 kHz and 16 bits word length linear PCM and can be played back by a commercial CD player. All speech samples in the data area are recorded in standardized format in 16-bit, 2's complement, low-byte first (little endian) format and can be retrieved by an ordinary PC-DOS system and CD-ROM reader.

C6.3.2 Recording Your Own Speech Databases

All speech recordings should be made in acoustical and electrical environments complying with the requirements given in Annex B.1.1 of ITU-T Rec. P.800.

The recommended method is to record the speech with a linear microphone and a low-noise amplifier with flat frequency response, digitize the speech, and then flat filter and level equalize. To achieve optimum SNR, the microphone should be positioned 15 to 20 cm from the talker's lips. A windscreen should be used if breath puffs from the talker are noticed.

The recordings should be made directly into a computer (A/D) or via a high quality recording system such as a DAT.

C6.3.3 Format for Single Sentence Speech Samples

Each source speech file will contain one sentence and will last nominally 4s. All source speech files within an experiment will be exactly the same length. This enhances the ability to recognize processing problems. An approximate 0.5 seconds period of silence precedes the sentence, and a similar period of silence follows the sentence. The speech files are organized as in the example shown in Figure 6.3.1. The sentences will be simple meaningful sentences as described in Annex B1.4 of ITU-T Rec. P.800.

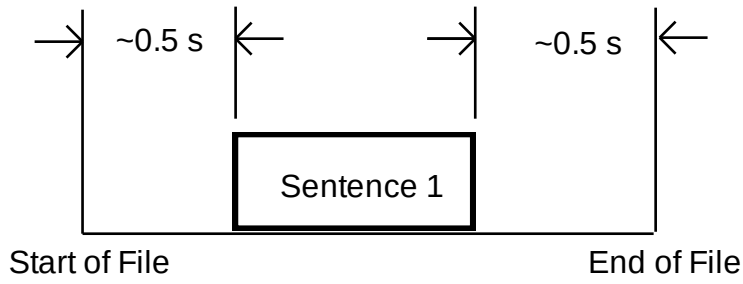

Figure 6.3.1: Example of Speech file structure for single sentences

It must be noted that the trailing silence of 0.5s after the end of the sentence in the file is of extreme importance, since there are (for some conditions) a series of FIR filters with large number of coefficients. If the prescribed trailing silence is not present, there is a considerable risk that speech will be clipped at the end of the file.

C6.3.4 Format for Short Speech Samples

Each source speech file will contain one pair of sentences and will last nominally 8 seconds, with a flexible time interval between the two sentences. All source speech files within an experiment will be

exactly the same length. This enhances the ability to recognize processing problems. An approximate 0.5 seconds period of silence precedes the first sentence in the file, and a similar period of silence follows the second sentence in the file. The speech files are organized as in the example shown in Figure 6.3.2. The sentences will be simple meaningful sentences as described in Annex B1.4 of ITU-T Rec. P.800.

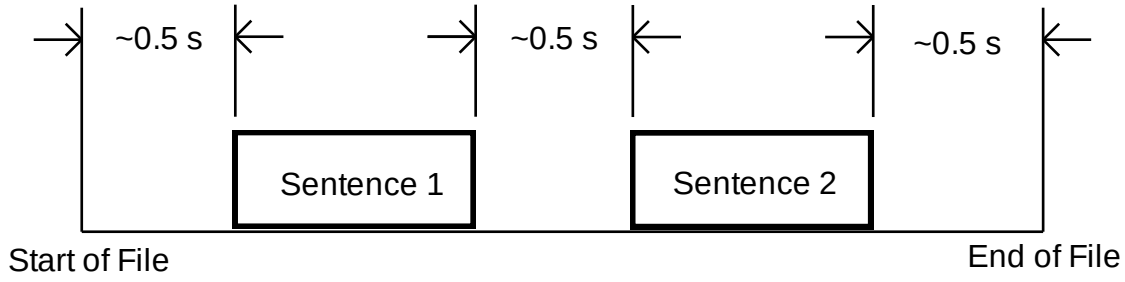

Figure 6.3.2: Example of speech file structure for short speech samples

It must be noted that the trailing silence of 0.5s after the end of the second sentence in the file is of extreme importance, since there are (for some conditions) a series of FIR filters with large number of coefficients. If the prescribed trailing silence is not present, there is a considerable risk that speech will be clipped at the end of the file.

C6.3.5 Format for Long Speech Samples

Each sample will contain 4 different sentences and will last nominally 16 seconds, with a time interval between sentences as described in Annex B1.4 of ITU-T Rec. P.800. All source speech files within an experiment will be exactly the same length. An approximate 0.3-0.5 seconds period of silence precedes the first sentence in the file, and a similar period of silence follows the last sentence in the file. The speech files are organized as in the example shown in Figure 6.3.3. The sentences will be simple meaningful sentences as described in Annex B1.4 of ITU-T Rec. P.800. Active speech in each source speech file should be present for not less than 9 seconds and not more than 12s. *{note – this last requirement may be hard to meet for some speech data bases. The typical English Harvard Sentence is less that 2 seconds long. Four of these would be less than the required 9 seconds of active speech. Therefore a reasonable relaxation of this last requirement should be tolerated}.*

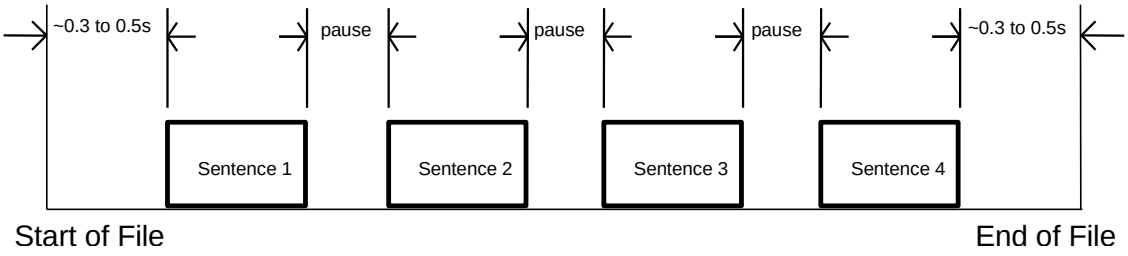

Figure 6.3.3: Example of speech file structure for long speech samples

These samples could be built by the addition of two of the 8-sec sentences described in Section 6.3.3, providing that the constraint for the active speech described above is (reasonably) fulfilled.

C6.3.6 Processing of the Speech Files

All speech files will need to be pre-processed prior to being processed through the experimental conditions. This pre-processing ensures that the speech is at the correct level and has the correct input characteristic. Full details on the processing required are given in [2]. Speech levels will be measured with the P.56 algorithm and level adjusted with the gain/loss algorithm to the level required for each test condition as defined in the test plans for the individual experiments. Where

the nominal level is specified, this level should be set to 26dB (±1dB) below digital overload (-26dBovl).

Some of the experiments require that the source speech material has background noise added. Details of the process to be followed are given in [2]. Noise levels will be measured with the rms. computation algorithm and level adjusted with the gain/loss algorithm to the required level. The following procedure will be followed:

i. The environmental noise will be Delta SM filtered to incorporate a near field microphone response.

ii. The environmental noises will be passed through the GSM send characteristic (see [2]).

iii. The noise levels will be adjusted using the r.m.s. measure to the mean level dictated by following test plans. For each type of noise, six segments will be taken from the noise file. The segments will be numbered from N1 to N6.

iv. The source speech material will be passed through the GSM send characteristic [2] and normalized (level equalized to -26dB) using the speech level meter complying with Rec. P.56. This is the responsibility of the Host Laboratories.

v. Finally, the noise will be digitally mixed with the normalized speech material. If the resulting signal amplitude exceeds the overload point of the A/D converter, it should be limited to the peak value and the clipping effect should be controlled by expert observation. The following mixing scheme details the combining of speech and noise samples for each speaker.

	M1	**M2**	**F1**	**F2**
Speech sample 1	N1	N2	N3	N4
Speech sample 2	N2	N3	N4	N5
Speech sample 3	N3	N4	N5	N6
Speech sample 4	N4	N5	N6	N1
Speech sample 5	N5	N6	N1	N2
Speech sample 6	N6	N1	N2	N3
Speech sample 7 (practice)	N1	N3	N5	N2

Table 6.3.4: Speech vs. Noise samples mixing scheme

C6.4 Listening Environment

For all experiments, subjects should be seated in a quiet environment; 30dBA Hoth Spectrum (as defined by ITU-T, Recommendation P.800, Annex A, section A.1.1.2.2.1 Room Noise, with table A.1 and Figure A.1) measured at the head position of the subject. This will help ensure consistency between the different subjects in the same laboratory as well as across the different laboratories in which these experiments will be performed.

The following points should be adhered to:

- Where the experiment design and the listening environment allows for multiple subjects in each listening session, the requirements stated above apply to each of the positions the subjects will occupy.
- Where there are multiple simultaneous subjects, they should not be able to see the responses made by other subjects.

- All test stimuli will be presented to the subjects over a telephone handset with Modified IRS receiving response (exclusive of the SRAEN filter). Any deviation shall be reported, e.g. use of one ear-piece in a headphone.
- Subjects should be told not to discuss the experiment with subjects who are yet to participate.
- Any test house performing multiple experiments must use different listening subjects for each experiment or sub-experiment.

C6.5 Experimental Procedure

Initially the experimenter should present and explain the experiment instructions to the subjects. When the subject has understood the instructions, they will first listen and give score to the preliminary conditions. After the preliminaries have been completed, there should be sufficient time allowed for answering possible questions from the subjects. Any questions about the procedure or the meaning of the instructions should be answered, but any technical questions on matters such as the experimental methodology or details of the types of distortions they are listening to must not be answered until they have completed the experiment.

C6.6 Preliminary Conditions

Preliminary conditions are included in the experiment to help acclimatize the subjects with the experimental procedure and to help reduce learning effects of the subjects, by ensuring that the subjects hear a full range of the potential qualities at the start of the experiment. No suggestions should be made to the subjects that the preliminary samples include the best or worst in the range to be covered, or exhaust the range of conditions they can expect to hear.

C6.7 Reference Conditions

Four types of reference conditions are used in these experiments:

- AMR without NS References: These are to be used to determine how the AMR Noise Suppresser performs in relation to these.
- Direct unprocessed speech plus noise source material.
- MNRU references: These are included as standard references of known and well understood performance and will allow the results to be expressed in terms of Equivalent Q as well as MOS for the ACR tests. MNRUs are also included in the CCR tests as references to estimate the test sensitivity and explore most of the CMOS range. For the CCR experiments, relative MNRU comparisons are used to estimate the test sensitivity. For example an MNRU of 12 may be compared to an MNRU of 16. If this difference is just above the significance level, it represents the test sensitivity.
- Ideal noise suppression levels: These are represented by varying the SNR level between the speech and noise. These conditions are AMR processed. This is an attempt to define equivalent noise suppression levels.

For the Tests involving background noise conditions, the MNRU references will use noisy speech (i.e. background noise will be used with the MNRU). The exact number of each of these types of reference in each experiment can be found in the experiment plans in the sections 7-10.

C6.8 Noise Material

Most of the Noise Suppresser Selection Test Experiments require the addition of noise to the speech material. The following types of noise are identified in this test plan:

Car Noise: This represents stationary (static) background noise and will be typical of the noise experienced when inside a moving vehicle (car) at a constant speed.

Street Noise: This represents non-stationary (dynamic) noise and will be typical of noise which might be experienced by someone using a mobile on a city street.

Babble Noise: This represents non-stationary (dynamic) noise and will be typical of the background noise encountered in public places: restaurant, cafeteria, open offices.

Noise files available free of charge from ARCON solely for the purposes of SMG11/S4 work shall be used. Contact ETSI (Paolo Usai) for further information (Paolo.Usai@ETSI.FR)

C7. Experiment 1: Degradation in Clean Speech (Pair Comparison Test)

C7.1 Introduction

This PC (Paired-Comparison) experiment was prepared to test the '**No degradation in clean speech**' requirement in the Recommended Minimum Performance Requirements specification ([1], TS GSM 06.77), i.e.. 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."

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.

C7.2. 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.

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 (Table 7.1) shows the testing factors to be used in this experiment. Due to the limited number of conditions tested within this experiment, it is possible to design a more balanced test structure and introduce some dummy conditions where the perceived difference in quality within the pairs of stimuli should be obvious for the subjects. A list of test conditions is given in Table 7.3.

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	0	No tandeming condition
Input Characteristic	1	GSM Filtered
Codec references	**#**	***Notes***
Test vocoders	1	AMR with NS
Reference vocoder	8	AMR at 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15 & 4.75
Other references	#	Notes
Direct		Nominal level, GSM Filtered
MNRU	2	Q = 5 dB & 20 dB, other Q values 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 7.1: Factors and conditions for Experiment 1

C7.3 Preliminary Conditions

The following 16 preliminary test conditions are recommended.

Cond.	Presentation order	Reference Codec	Trans-codings	Processed Codec	Trans-codings	Talker and Sample Number
P1	5	Direct	-	MNRU-20	-	F1S13
P2	1	MNRU-18	-	MNRU-22	-	M1S13
P3	3	MNRU-19	-	MNRU-21	-	F2S13
P4	7	AMR-12.2	1	AMR-12.2	1	M2S13
P5	6	AMR-12.2	1	AMR-5.9	1	F1S13
P6	2	AMR-5.9	1	AMR-5.9	1	M1S13
P7	4	AMR-4.75	1	AMR-7.95	1	F2S13
P8	8	MNRU-5	-	MNRU-20	-	M2S13
P9	14	MNRU-20	-	Direct	-	F1S13
P10	10	MNRU-22	-	MNRU-18	-	M1S13
P11	12	MNRU-21	-	MNRU-19	-	F2S13
P12	16	AMR-12.2	1	AMR-12.2	1	M2S13
P13	13	AMR-5.9	1	AMR-12.2	1	F1S13
P14	9	AMR-5.9	1	AMR-5.9	1	M1S13
P15	11	AMR-7.95	1	AMR-4.75	1	F2S13
P16	15	MNRU-20	-	MNRU-5	-	M2S13

Table 7.2: List of preliminary conditions for Experiment 1

C7.4 Speech Material

Single sentences. For the 4 talkers, 2 male and 2 female there are:

13 stimuli / talker, each stimuli 4sec long w/ 1 sentence

12 unique sentences / talker for test plus one for practice

To reduce the speech material effect, each talkers' samples must be unique. For this experiment, the unique samples are not balanced across all condition, candidates and subject groups. The same sample numbers for each talker are used for common conditions within a subject group and changed across subject groups.

C7.5 Experimental Design

The design is based on a restricted randomization philosophy using 6 different randomizations, each one covered by a group of 4 of the 24 subjects. This means that up to 4 subjects can perform the experiment simultaneously.

Each subject will hear all of the conditions 16 times, four times with speech from each of the four talkers. Each of two stimuli for a talker will be presented in both the A/B and B/A order. Over the experiment as a whole, each of the conditions will be paired with twelve different samples from each of the four talkers. Each of the six groups of subjects will hear different combinations of source material and condition.

C7.6 Processing

Every condition has to be processed for each of the twelve stimuli of each of the four talkers. The actual samples used for each condition by each subject group are presented in Section 7.12 Test Conditions.

C7.7 Randomizations

Separate randomizations for each of the six subject groups shall be provided to reduce order effects and to minimize differences between the laboratories. There shall be six randomizations for the experiment, one for each subject group. Each one will therefore be used by four of the 24 subjects.

C7.8 Duration of the PC Experiment

Each stimuli is 4 sec reference + 4 sec speech sample + 4 s voting time or 12 seconds. For this experiment there are 16 preliminary conditions x 12 seconds or 3.2 minutes for an introductory block. The presentation set for the experiment consists of 40 conditions (A/B+B/A) x 2 repeats x 4 talkers x 12 seconds or 64 minutes. The experiment is presented as the 16 preliminary conditions followed by the test itself divided in several sessions, i.e. 67,2 minutes testing time / subject group. The 6 groups of 4 subjects require 7 hours and 30 minutes total testing time for the experiment (6 x 1h 15 env.)

To reduce the effects of subject fatigue, sessions should be separated by short comfort breaks.

Note that the above calculations do not include the time needed to give the subjects their instructions, or for comfort breaks.

C7.9 Votes Per Condition

Every condition will have 24 subjects vote on four stimulus from each of four talkers, giving:

(24 subjects x 4 talkers x 4 Presentations) = 384 votes per condition

From past experience of PC tests, this is the minimum number of votes per condition needed to give enough statistical certainty to differentiate the performance of one candidate process from another candidate process over the conditions and against the references.

C7.10 Test Procedure

Factors important for the experimental environment are specified in sections 6.4, 6.5, and 6.6. As specified in section 7.8, comfort breaks should be provided to reduce the effects of subject fatigue.

7.11 Opinion ScaleThe question asked of the subject is according to the Paired-Comparison binary scale. The specific wording is designed to evaluate the relative quality of the test sample in relation to the reference sample. In order to minimise presentation bias, the samples will be presented in both the A/B and B/A directions within the experiment. The subjects will listen to each pair of samples, and after presentation is completed, they will be asked to give their opinion. Annex A.1 contains an example of the instructions for the subjects in English.

C7.12. 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):

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

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

C7.13. 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@10.2	AMR@10.2	1	3 4 5 6 1 2
3	AMR@7.95	AMR@7.95	1	1 2 3 4 5 6
4	AMR@7.4	AMR@7.4	1	4 5 6 1 2 3
5	AMR@6.7	AMR@6.7	1	5 6 1 2 3 4
6	AMR@5.9	AMR@5.9	1	6 1 2 3 4 5
7	AMR@5.15	AMR@5.15	1	2 3 4 5 6 1
8	AMR@4.75	AMR@4.75	1	3 4 5 6 1 2
9	AMR@12.2	AMR@5.9	1	1 2 3 4 5 6
10	AMR@4.75	AMR@7.95	1	4 5 6 1 2 3
11	DIRECT	MNRU Q= 20 dB	1	5 6 1 2 3 4
12	MNRU Q= 5 dB	MNRU Q= 20 dB	1	6 1 2 3 4 5
13	AMR@12.2	AMR/NS@12.2	1	2 3 4 5 6 1
14	AMR@10.2	AMR/NS@10.2	1	3 4 5 6 1 2
15	AMR@7.95	AMR/NS@7.95	1	1 2 3 4 5 6
16	AMR@7.4	AMR/NS@7.4	1	4 5 6 1 2 3
17	AMR@6.7	AMR/NS@6.7	1	5 6 1 2 3 4
18	AMR@5.9	AMR/NS@5.9	1	6 1 2 3 4 5
19	AMR@5.15	AMR/NS@5.15	1	1 2 3 4 5 6
20	AMR@4.75	AMR/NS@4.75	1	3 4 5 6 1 2
21 – 40	Reversed order of the reference and processed speech samples in cond. 1-20			
41 – 60	Repeat of conditions 1 – 20 with Speech Sample Number +6			
61 - 80	Reversed order of the reference and processed speech samples in cond. 41 - 60			
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 7.3: Test conditions for Experiment 1

C8 Experiments 2a, 2b & 2c: No degradation of Speech and no Undesirable Effects in Residual Noise in Conditions with Background Noise (ACR)

C8.1 Introduction

These ACR experiments are designed to test the requirement "No degradation of Speech and no Undesirable Effects in Residual Noise" in the Minimum Performance Requirements for Noise Suppresser Application to the AMR Speech Encoder, [1],. These ACR experiments will be run for three types of acoustic background noise.

C8.2 Test Factors and Conditions

The ACR 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 ACR experiments with the different noise types in separate experiments. Within each experiment, a low, a medium 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 and SNR + 12 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).

All three experiments are run at AMR bit rate 12.2 kbit/s and 5.9 kbit/s.

The following table shows the testing factors to be used in these experiments. A full list of test conditions is given in Section 8.12.

Main Codec Conditions	**#**	**Notes**
Noise Suppresser Algorithms	1	
Codec	1	AMR
Codec Modes	2	12.2 kbps rate, 5.9 kbps rate
BERs	0	Clear channel, no transmission errors
Input level	3	nominal (high, low): -26dB (-16 dB, -36 dB) relative to OVL
Acoustic Background Noise	3	Static Car @ 6dB, 12dB, 18dB Street @ 9dB, 15dB, 21dB Babble @ 9dB, 15dB, 21dB
Input Characteristic	1	GSM Filtered
VAD/CNG/DTX	2	ON only at the nominal level, medium SNR values, zero value of Ideal NS OFF for other conditions One VAD/CNG/DTX will be used ; either VAD Option 1 or 2, depending on the implementers choice
Codec references	**#**	**Notes**
All Experiments	1	AMR wo/ NS
Other references	**#**	**Notes**
Direct		Nominal level, GSM Filtered
MNRU, Exp 2a, 2b, 2c	5	Nominal level, with background noise, GSM Filtered, Q= 6, 12, 18, 24, 30dB
Ideal Noise Suppression	6	3 levels for each SNR
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 ACR Instructions
Replications	1	Original Presentation Only

Table 8.2.1: Factors and conditions for Experiments 2a, 2b, 2c

C8.3 Preliminary Conditions

The following 16 preliminary test conditions are recommended.

Cond.	Presentation order	SNR value	Ideal NS (dB)	Codec	Talker and Sample Number
P1	5	SNR	-	Direct	M1S07
P2	1	SNR	-	MNRU-12	M2S07
P3	3	SNR	-	AMR@12.2	M1S07
P4	7	SNR	7	AMR@12.2	M2S07
P5	6	SNR+6	7	AMR@12.2	F1S07
P6	2	SNR+12	7	AMR@12.2	F2S07
P7	4	SNR	-	AMR@5.9	F1S07
P8	8	SNR+12	-	AMR@5.9	F2S07
P9	14	SNR	-	Direct	F1S07
P10	10	SNR	-	MNRU-12	F2S07
P11	12	SNR	-	AMR@12.2	F1S07
P12	16	SNR	7	AMR@12.2	F2S07
P13	13	SNR+6	7	AMR@12.2	M1S07
P14	9	SNR+12	7	AMR@12.2	M2S07
P15	11	SNR	-	AMR@5.9	M1S07
P16	15	SNR+12	-	AMR@5.9	M2S07

Table 8.3.1: List of preliminary conditions

C8.4 Speech Material

The speech material should be as defined in Section 6.4 - Long Sentence Quads, with each sample containing 4 sentences. For each test condition there are:

6 samples / talker, each sample 16sec long w/ 4 sentences

24 unique sentences / talker

For the practice conditions there are:

1 sample / talker

4 unique sentences / talker

To reduce any speech material effect, each talker sample must be unique. For these experiments, the unique samples are not balanced across all condition, candidates and subject groups. The same sample numbers for each talker are used for common conditions within a subject group and changed across subject groups. For a given language, the same speech material must be used for the three experiments 2a, 2b and 2c.

Speech samples numbered from 01 to 06 should be used for the test conditions; speech samples numbered as 07 should be used for the Practice session.

The noise material and its mix with the speech material should be as defined in Section 6.10 and Section 8.2.

C8.5 Experimental Design

The design is based on a restricted randomization philosophy using 6 different randomizations, each one covered by a group of 4 of the 24 subjects. This means that up to 4 subjects can perform the experiment simultaneously.

Each subject will hear all of the conditions four times, once with speech from each of the four talkers. Over the experiment as a whole, each of the conditions will be paired with six different samples from each of the four talkers. Each of the six groups of subjects will hear different combinations of source material and condition.

C8.6 Processing

Every condition has to be processed for each of the six stimuli of each of the four primary talkers. The actual samples used for each condition by each subject group are presented in Section 8.12 Test Conditions.

C8.7 Randomizations

Separate randomizations for each of the six subject groups shall be provided to reduce order effects and to minimize differences between the laboratories. There shall be six randomizations for the sub-experiments, one for each subject group. The same randomizations will be used for the three experiments (2a, 2b and 2c). Each one will therefore be used by four of the 24 subjects. Each randomization shall be balanced across 4 blocks of 36 stimuli to eliminate long sequences of similar conditions or identical talkers. The sequences shall provide for alternating male-female talkers.

C8.8 Duration of the ACR Experiments 2a, 2b, and 2c

Each stimuli is 16 s speech sample + 5 s voting time or 21 seconds. For each of the three experiments there are 16 preliminary conditions x 21 seconds or 5.6 minutes for an introductory block. The test consists of 36 conditions x 4 talkers x 21 seconds or 50.4 minutes, presented as three 16.8 minute blocks of 36 stimuli for 56 minutes testing time / subject group. The 6 groups of 4 subjects require 4 hours and 24 minutes total testing time

To reduce the effects of subject fatigue, the three blocks should be separated by short comfort breaks.

Note that the above calculations do not include the time needed to give the subjects their instructions, or for comfort breaks.

C8.9 Votes Per Condition

In each of the three experiments, every condition will have 24 subjects vote on one stimulus from each of four talkers, giving:

(24 subjects x 4 talkers) = 96 votes per condition

From past experience of ACR tests, this is the minimum number of votes per condition needed to give enough statistical certainty to differentiate the performance of one candidate process from another candidate process over the conditions and against the references.

C8.10 Test Procedure

Factors important for the experimental environment are specified in section 6.5 and 6.6. As specified in section 9.8, comfort breaks should be provided to reduce the effects of subject fatigue.

C8.11 Opinion Scale

The question asked of the subject is a modification of the ACR Listening Quality Scale. The specific wording is designed to evaluate both the level of distortion of the speech and the presence of artifacts in the residual background noise signal. The subjects will listen to each sample and after it has completed they will be asked to give their opinion.

Annex A contains an example of the instructions for the subjects in English. The instructions in Annex A contain a modified version of the ACR instructions. They are aimed at focusing the subjects to rate artifacts introduced by the NS device. The test administrator should have the freedom to provide guidance to the subjects to reinforce this point, provided that such instructions are consistent across all 24 subjects. This is particularly important for tests not performed in English. Any additional instructions given to the subjects should be reported as an integral part of test reports.

C8.12 Test Conditions for Experiments 2a, 2b and 2c

Cond.	Input level	SNR value	Ideal NS (dB)	VAD/DTX	Codec	Speech sample number (6 sequences)
1	nominal	SNR	-	N/A	Direct	4 5 6 1 2 3
2	nominal	SNR	-	N/A	MNRU-30	4 5 6 1 2 3
3	nominal	SNR	-	N/A	MNRU-24	4 5 6 1 2 3
4	nominal	SNR	-	N/A	MNRU-18	4 5 6 1 2 3
5	nominal	SNR	-	N/A	MNRU-12	4 5 6 1 2 3
6	nominal	SNR	-	N/A	MNRU-6	4 5 6 1 2 3
7	nominal	SNR	-	off	AMR@12.2	1 2 3 4 5 6
8	nominal	SNR	4	off	AMR@12.2	1 2 3 4 5 6
9	nominal	SNR	7	off	AMR@12.2	1 2 3 4 5 6
10	nominal	SNR	-	off	AMR@5.9	1 2 3 4 5 6
11	high	SNR	-	off	AMR@12.2	1 2 3 4 5 6
12	high	SNR	-	off	AMR@5.9	1 2 3 4 5 6
13	nominal	SNR+6	-	off	AMR@12.2	2 3 4 5 6 1
14	nominal	SNR+6	4	off	AMR@12.2	2 3 4 5 6 1
15	nominal	SNR+6	7	off	AMR@12.2	2 3 4 5 6 1
16	nominal	SNR+6	-	off	AMR@5.9	2 3 4 5 6 1
17	nominal	SNR+6	-	on	AMR@12.2	2 3 4 5 6 1
18	nominal	SNR+6	-	on	AMR@5.9	2 3 4 5 6 1
19	low	SNR+6	-	off	AMR@12.2	2 3 4 5 6 1
20	low	SNR+6	-	off	AMR@5.9	2 3 4 5 6 1
21	nominal	SNR+12	-	off	AMR@12.2	3 4 5 6 1 2
22	nominal	SNR+12	4	off	AMR@12.2	3 4 5 6 1 2
23	nominal	SNR+12	7	off	AMR@12.2	3 4 5 6 1 2
24	nominal	SNR+12	-	off	AMR@5.9	3 4 5 6 1 2
25	nominal	SNR	-	off	AMR/NS@12.2	1 2 3 4 5 6
26	nominal	SNR	-	off	AMR/NS@5.9	1 2 3 4 5 6
27	nominal	SNR+6	-	off	AMR/NS@12.2	2 3 4 5 6 1
28	nominal	SNR+6	-	off	AMR/NS@5.9	2 3 4 5 6 1
29	nominal	SNR+12	-	off	AMR/NS@12.2	3 4 5 6 1 2
30	nominal	SNR+12	-	off	AMR/NS@5.9	3 4 5 6 1 2
31	nominal	SNR+6	-	on	AMR/NS@12.2	2 3 4 5 6 1
32	nominal	SNR+6	-	on	AMR/NS@5.9	2 3 4 5 6 1

33	low	SNR+6	-	off	AMR/NS@12.2	2 3 4 5 6 1
34	low	SNR+6	-	off	AMR/NS@5.9	2 3 4 5 6 1
35	high	SNR	-	off	AMR/NS@12.2	1 2 3 4 5 6
36	high	SNR	-	off	AMR/NS@5.9	1 2 3 4 5 6

Note: 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

C8.13 Statistical Analysis

The statistics to be reported from this ACR test are the averaged MOS (MOS_k) scores and the standard deviations (S_k) for all the conditions.

Additionally, the requirement in [1, Section 6.1.3] should be checked using a hypothesis test for the conditions 25-36 if the mean MOS score is greater or equal to the MOS score for the corresponding equivalent (all being equal except NS activated) reference condition for AMR without NS within a 95 % confidence.

The hypothesis test should be performed using a 2-tailed T-test. The NS algorithm has failed the requirement if, for any of test condition,

$$t < -t_{N,0.05}$$

where

$$t = \frac{\mathrm{MOS}_{test} - \mathrm{MOS}_{ref}}{\sqrt{\frac{\mathrm{S}_{test}^2 + \mathrm{S}_{ref}^2}{N}}}$$

and the subscripts $_{test}$ and $_{ref}$ denotes the test condition and corresponding reference condition, respectively, N is the number of votes, and $t_{N,0.05}$ is the inverse of the Student's t-distribution with N degrees of freedom and probability 0.05.

C9. Experiments 3a & 3b: Performances in Background Noise Conditions (Mod-CCR)

C9.1 Introduction

These experiments are designed to test Requirements in the associated Section in the Recommended Minimum Performance Requirements Specification ([1], TS GSM 06.77). Specifically, the AMR with noise suppression should, in a certain number of conditions, be preferred to the AMR without noise suppression in a background noise environment and should provide a reasonable level of SNR improvement. Experiment 3a examines the performance of the noise suppression with the half-rate codec, while Experiment 3b examines the noise suppression with the full rate codec. Both experiments will use the Modified Comparison Category Rating (Mod-CCR, Note 1) method with a seven-point rating scale. Listeners will judge the relative quality of samples processed through the codec with noise suppression, compared to those without the noise suppression applied (example instructions for listeners are given in Annex A.3). The samples will have background noise of various types and levels mixed into the source speech before processing through the codec.

The factors for each of the four sub-experiments are presented in Table 9.1.

Factor	*Expt 3a*	*Expt 3b*
codec	AMR 5.9 kb/s	AMR 12.2 kb/s
noise types	car (6 and 15 dB) street (9 and 18 dB) babble (9 and 18 dB)	car (6 and 15 dB) street (9 and 18 dB) babble (9 and 18 dB)

Table 9.1: Factors for Experiments 3a and 3b

Note 1:

The standard Comparison Category Rating method (CCR) which is described in Annex E of Rec. P.800 is similar to the Degradation Category Rating method (DCR, Annex D). In Annex E, it is explicitly said : "Listeners are presented with a pair of speech samples on each trial. In the DCR procedure, a reference (unprocessed) sample is presented **first**, followed by the same speech sample, which has been processed by some technique. In the CCR procedure, the order of the processed and unprocessed samples is chosen at random for each trial. Listeners use the seven-point CCR scale to judge the quality of the second sample relative to that of the first. **The DCR and the CCR methods are particularly useful for assessing the performance of telecommunications systems when the input has been corrupted by background noise. However, an advantage of the CCR method over the DCR procedure is the possibility to assess speech processing that either degrades or improves the quality of the speech.**

Here we are using a different application of the standard CCR method. The modified CCR method uses processed reference samples (but without noise suppression applied) whereas the standard CCR method uses unprocessed reference samples.

C9.2 Test Factors and Conditions

Three types of background noise will be used, at two different SNRs:

- 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.

The noise samples will be those utilised during the AMR Noise Suppression Selection Phase.

The codec is held constant for each experiment, with two SNR classes ('SNR' and 'SNR+9dB') per experiment. All of the noise types are used in each experiment. The noise samples will have been recorded in scenarios representative of the respective SNR value for each noise.

The factors and conditions to be used in Experiments 3a and 3b are presented in Table 9.2. The expanded set of test conditions is given in Section 9.12.

Main Codec Conditions	**#**	**Notes**
Noise Suppresser Candidates	1	
Codec	1	AMR
Codec Modes (HR/FR)	HR FR	5.9 kbit/s rate for Experiment 3a 12.2 kbps rate for Experiment 3b
BERs	0	Clear channel, no transmission errors
Input level	1	nominal: -26dB relative to OVL
Acoustic Background Noise	3	car, street, and babble noise
Background noise SNRs	2	low, high for each (see Table 9.1)
Input Characteristic	1	GSM transmit filtered
Codec references	**#**	**Notes**
All Experiments	1	**the same AMR rate w/o NS**
Other references	**#**	**Notes**
Direct		nominal level, GSM transmit filtered
MNRU, Exp 3a and 3b		nominal level, GSM transmit filtered, Q= 12, ΔQ= 4
Ideal noise suppression simulation		
Common Conditions	**#**	**Notes**
GSM Channel	0	NO channel model
Number of talkers	4	2 male + 2 female primary talkers
Number of speech samples	28	7 Sentence-pairs/primary talker (6 for Test, 1 for Practice)
Listening Level	1	-15dBPa (79dB SPL) at ERP
Listeners	24	**Naive Listeners**
Randomizations	6	6 groups of 4 listeners
Rating Scale	1	CCR Instructions
Replications	1	Original Presentation Only

Table 9.2: Factors and conditions for Experiments 3a and 3b

C9.3 Preliminary Conditions

The following 16 preliminary test conditions are recommended, for presentation, before proceeding to the test samples. The samples shall be presented in the random order given in Table 9.3

Cond.	Presentation order	Noise	SNR (dB)	Reference	Processed		Speech Sample Number
					Ideal NS	Codec	
P1	9	Car	6	Direct	-	Direct	M1S07
P2	5	Car	15	AMR@x	-	AMR@x	F1S07
P3	12	Car	6	MNRU-12	-	MNRU-16	M2S07
P4	13	Car	15	MNRU-12	-	Direct	F2S07
P5	2	Street	9	AMR@x	-	AMR@x	M1S07
P6	4	Street	18	MNRU-12	-	MNRU-16	F1S07
P7	8	Street	18	MNRU-12	-	Direct	M2S07
P8	16	Babble	9	AMR@x	-	AMR@x	F2S07
P9	7	Babble	9	MNRU-12	-	MNRU-16	M1S07
P10	1	Babble	18	MNRU-12	-	Direct	F1S07
P11	11	Car	6	AMR@x	4	AMR@x	M2S07
P12	3	Car	15	AMR@x	10	AMR@x	F2S07
P13	15	Street	18	AMR@x	4	AMR@x	M1S07
P14	6	Street	9	AMR@x	10	AMR@x	F1S07
P15	10	Babble	9	AMR@x	4	AMR@x	M2S07
P16	14	Babble	18	AMR@x	10	AMR@x	F2S07
Notes:	- The bit rate for the AMR processing for the preliminary samples shall be the same as that used for the test samples, 5.9 kbit/s for Experiment 3a, 12.2 kbit/s for Experiment 3b.						

Table 9.3: List of preliminary conditions

C9.4 Speech Material

The source speech material shall be as defined in Section 6.3 and will consist of the material used during the AMR Noise Suppression Selection phase: Each sample consists of two sentences. Only primary talkers are needed. For the four talkers, the following source material should be prepared:

Seven samples for each talker, six for the test samples and one for the preliminaries,

Each sample to be eight seconds long,

Unique sentences-pairs in each sample (i.e., no repeated across the talkers)

To reduce any speech material effect, the samples for each talker must be unique. For these experiments, these unique stimuli are not balanced across all conditions, candidates and subject groups. The same sample numbers for each talker are used for common conditions within a subject

group and changed across subject groups (these sample numbers are arbitrarily assigned to samples). For a given language, the same speech material must be used for the two experiments 3a and 3b. The noise material and its mix with the speech material should be as defined in Section 6.8 and Section 6.3.7 respectively.

C9.5 Experimental Design

The design is based on a restricted randomization philosophy using six different randomizations, each of which is used with a group of four of the 24 listeners. This means that up to four subjects can perform the experiment simultaneously.

Each listener will hear all of the conditions four times, once with speech from each of the four talkers. Over the experiment as a whole, each of the conditions will be paired with six different samples from each of the four talkers. Each of the six groups of subjects will hear different combinations of source material and condition.

C9.6 Processing

Every condition is processed with each of the six samples of each of the four primary talkers. The actual samples to be used for each condition, within with each subject group, are presented in Section 9.12, *Test Conditions*.

C9.7 Randomizations

The test shall be completed using the randomizations provided by the experimenter. There shall be six randomizations for the sub-experiments, one for each subject group. The same randomizations shall be used for the two experiments (3a and 3b). Each one will therefore be used by four of the 24 subjects. Each randomization is balanced across four blocks of 48 stimuli to eliminate long sequences of similar conditions or identical talkers. The sequences shall provide for alternating male-female talkers. Use of these randomizations will allow presentation order to be used as a factor in a global analysis, should that be necessary. The randomization shall be constrained to a randomized block design, which controls practice and fatigue effects that may occur over the course of a test session.

C9.8 Duration of the CCR Experiments 3a and 3b

Each trial consists of an eight-second reference sample + an eight-second test sample + five second voting time, totaling 21 seconds. For each of the four experiments there are 16 preliminary conditions x 21 seconds or 5.6 minutes for an introductory block. Each presentation set within an experiment consists of 52 conditions (A/B+B/A) x 4 talkers x 21 seconds or 70 minutes, presented as eight 8.75 minute blocks of 25 stimuli for 75.6 minutes testing time / subject group / experiment. The total testing time for each experiment will be 7 hours and 34 minutes, if four listeners are tested at one time.

To reduce the effects of subject fatigue, each 8.75 minute block should be separated by short comfort breaks.

Note that the above calculations do not include the time needed to give the subjects their instructions, or time taken for comfort breaks.

C9.9 Votes Per Condition

In each of the three experiments, 24 listeners rate every condition with four talkers in each of two presentation orders (A/B and B/A), giving:

(24 subjects x 4 talkers x 2 presentations) = 192 votes per condition

From past experience with CCR tests, this is the minimum number of votes per condition needed to give enough statistical certainty to differentiate the performance of one candidate process from another candidate process over the conditions and against the references.

C9.10 Test Procedure

Factors important for the experimental environment are specified in Sections 6.4, 6.5, and 6.6. As specified in Section 9.8, comfort breaks should be provided to reduce the effects of subject fatigue.

C9.11 Opinion Scale

The question asked of the subject is a based on of the CCR Listening Quality Comparison Scale. The listening subjects will judge the quality of the second sample with regard to quality of the first sample. The subjects will listen to each pair of samples and after these have been played, they will be asked to give their comparative opinion. Annex A contains an example of the instructions for the subjects in English. Changes to the instructions may be needed to specify the method of data collection being used (button-press, paper & pencil, etc.).

C9.12 Test Conditions for Experiments 3a and 3b

Cond.	Noise	SNR (dB)	Reference	Processed		Speech sample number
				Ideal NS	Codec	
1	Car	6	AMR@x	-	AMR@x	4 5 6 1 2 3
2	Street	9	AMR@x	-	AMR@x	4 5 6 1 2 3
3	Babble	9	AMR@x	-	AMR@x	4 5 6 1 2 3
4	Car	6	MNRU-16	-	MNRU-12	4 - - 1 - -
5	Car	6	Direct	-	MNRU-12	4 - - 1 - -
4'	Street	9	MNRU-16	-	MNRU-12	- 5 - - 2 -
5'	Street	9	Direct	-	MNRU-12	- 5 - - 2 -
4"	Babble	9	MNRU-16	-	MNRU-12	- - 6 - - 3
5"	Babble	9	Direct	-	MNRU-12	- - 6 - - 3
6	Car	6	AMR@x	3	AMR@x	1 2 3 4 5 6
7	Car	6	AMR@x	6	AMR@x	1 2 3 4 5 6
8	Car	6	AMR@x	9	AMR@x	1 2 3 4 5 6
9	Street	9	AMR@x	3	AMR@x	2 3 4 5 6 1
10	Street	9	AMR@x	6	AMR@x	2 3 4 5 6 1
11	Street	9	AMR@x	9	AMR@x	2 3 4 5 6 1
12	Babble	9	AMR@x	3	AMR@x	3 4 5 6 1 2
13	Babble	9	AMR@x	6	AMR@x	3 4 5 6 1 2
14	Babble	9	AMR@x	9	AMR@x	3 4 5 6 1 2
15	Car	6	AMR@x	-	AMR/NS1@x	1 2 3 4 5 6
16	Street	9	AMR@x	-	AMR/NS1@x	2 3 4 5 6 1
17	Babble	9	AMR@x	-	AMR/NS1@x	3 4 5 6 1 2
18	Car	15	AMR@x	3	AMR@x	1 2 3 4 5 6
19	Car	15	AMR@x	6	AMR@x	1 2 3 4 5 6
20	Street	18	AMR@x	3	AMR@x	2 3 4 5 6 1
21	Street	18	AMR@x	6	AMR@x	2 3 4 5 6 1
22	Babble	18	AMR@x	3	AMR@x	3 4 5 6 1 2
23	Babble	18	AMR@x	6	AMR@x	3 4 5 6 1 2
24	Car	15	AMR@x	-	AMR/NS1@x	1 2 3 4 5 6
25	Street	18	AMR@x	-	AMR/NS1@x	2 3 4 5 6 1

26	Babble	18	AMR@x	-	AMR/NS1@x	3 4 5 6 1 2
27-52	Reversed order of the reference and processed speech samples in cond. 1-26					
Notes:	- AMR@x denotes AMR at bit rate x, AMR/Ns1@x denotes the NS candidate at bit rate x; 5.9 kbit/s for Experiment 3a, 12.2 kbit/s for Experiment 3b - SNR(dB) denotes SNR for noise - 4 talkers are used for all conditions: 2 male and 2 female - 6 speech samples (8 s) are used for each talker - 'multiple' conditions "4s" and "5s" (e.g. 4 and 4') are only presented to a subset of listeners (e.g. to the first and the fourth groups of randomisation) ,					

C9.13 Statistical Analysis

The statistics to be reported from this CCR test are the averaged CMOS ($CMOS_k$) scores and the standard deviations (S_k) for all the conditions.

Additionally, the requirement in [1, Section 6.1.4] should be checked using hypothesis tests for the conditions 15-17 and 24-26 if the mean CMOS score is greater than zero (the NS performance is preferred) and greater or equal to zero (the NS performance is equivalent) within a 95 % confidence.

The hypothesis test should be performed using a 1-tailed T-test. The NS algorithm has failed the requirement at level "preferred" for any of test condition if

$$t < t_{N,0.05}$$

where

$$t = \frac{\mathrm{CMOS}_{test}}{\mathrm{S}_{test} / \sqrt{N}}$$

and the subscripts $_{test}$ denotes the test condition, N is the number of votes, and $t_{N,0.05}$ is the inverse of the Student's t-distribution with N degrees of freedom and probability 0.05.

Similarly, the NS algorithm has failed the requirement at level "equal" if

$$t < -t_{N,0.05}$$

C10 Experiments 4: Influence of Input Level, Voice Activity Detection and Discontinuous Transmission (CCR)

C10.1 Introduction

This experiment is designed to test Requirements in the associated Section in the Recommended Minimum Performance Requirements Specification ([1], TS GSM 06.77). Specifically, the AMR with noise suppression should, in a certain number of conditions, be preferred to the AMR without noise suppression in a background noise environment and should provide a reasonable level of SNR improvement.

C10.2 Test Factors and Conditions

Three types of background noise will be used, at two different SNRs:

- 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.

The factors and conditions to be used in Experiment 4 are presented in Table 10.2. The expanded set of test conditions is given in Section 10.12.

Main Codec Conditions	**#**	**Notes**
Noise Suppressor Candidates	1	NS algorithm under test
Codec	1	AMR
Codec Modes	1	12.2 kbit/s rate
BERs	0	Clear channel, no transmission errors
Input level	3	Nominal: -26dBov; High-level (-16 dBov); Low-level (-36 dBov)
Acoustic Background Noise	3	Car noise at 6 dB SNR; Street and Babble noise at 9 dB SNR
Input Characteristic	1	GSM transmit filtered
VAD/CNG/DTX	2	ON for all noise/level combinations OFF for all noise types but only at the nominal level One VAD/CNG/DTX will be used ; either VAD Option 1 or 2, depending on the implementers choice
Codec references	**#**	**Notes**
	1	**AMR 12.2 kbit/s rate without NS**
Other references	**#**	**Notes**
Direct	3	Nominal level, GSM transmit filtered
MNRU	6	Nominal level, GSM transmit filtered, Q=30, 24, 21, 18, 12, 6 dB, compared against Q=18 dB
Common Conditions	**#**	**Notes**
GSM Channel	0	NO channel model
Number of talkers	4	2 male + 2 female primary talkers
Number of speech samples	28	7 Sentence-pairs/primary talker (6 for Test, 1 for Practice)
Listening Level	1	-15dBPa (79dB SPL) at ERP
Listeners	24	**Naive Listeners**
Randomizations	6	6 groups of 4 listeners
Rating Scale	1	CCR Instructions
Replications	1	

Table 10.2: Factors and conditions for Experiment 4

C10.3 Preliminary Conditions

The following 16 preliminary test conditions are recommended, for presentation, before proceeding to the test samples. The samples shall be presented in the random order given in Table 10.3

Cond.	Presentation order	Noise	Input level	SNR (dB)	VAD/DTX	Reference	Processed	Speech Sample Number
P1	9	Car	nominal					M1S07
P2	5	Car	nominal	6		Direct	MNRU-12	F1S07
P3	12	Car	nominal	6		MNRU-16	MNRU-12	M2S07
P4	13	Car	nominal	15				F2S07
P5	2	Street	nominal	9		Direct	MNRU-12	M1S07
P6	4	Street	nominal	9		MNRU-16	MNRU-12	F1S07
P7	8	Street	nominal					M2S07
P8	16	Babble	nominal	9		Direct	MNRU-12	F2S07
P9	7	Babble	nominal	9		MNRU-16	MNRU-12	M1S07
P10	1	Babble	nominal					F1S07
P11	11	Car	nominal	6	off	AMR@12.2	AMR@12.2	M2S07
P12	3	Car	nominal	6	on	AMR@12.2	AMR@12.2	F2S07
P13	15	Street	nominal	9	off	AMR@12.2	AMR@12.2	M1S07
P14	6	Street	nominal	9	on	AMR@12.2	AMR@12.2	F1S07
P15	10	Babble	nominal	9	off	AMR@12.2	AMR@12.2	M2S07
P16	14	Babble	nominal	9	on	AMR@12.2	AMR@12.2	F2S07

Table 10.3: List of preliminary conditions [TO BE REVISED]

C10.4 Speech Material

The source speech material shall be as defined in Section 6.3 and will consist of the material used during the AMR Noise Suppression Selection phase: Each sample consists of two sentences. Only primary talkers are needed. For the four talkers, the following source material should be prepared:

Seven samples for each talker, six for the test samples and one for the preliminaries,

Each sample to be eight seconds long,

Unique sentences-pairs in each sample (i.e., no repeated across the talkers)

To reduce any speech material effect, the samples for each talker must be unique. For these experiments, these unique stimuli are balanced across all conditions, candidates and subject groups. The noise material and its mix with the speech material should be as defined in Section 6.8 and Section 6.3.7 respectively.

C10.5 Experimental Design

The design is based on a restricted randomization philosophy using six different randomizations, each of which is used with a group of four of the 24 listeners. This means that up to four subjects can perform the experiment simultaneously.

Each listener will hear all of the conditions four times, once with speech from each of the four talkers. Over the experiment as a whole, each of the conditions will be paired with six different samples from each of the four talkers. Each of the six groups of subjects will hear different combinations of source material and condition.

C10.6 Processing

Every condition is processed with each of the six samples of each of the four primary talkers. Every speech file will be processed through all test conditions.

C10.7 Randomizations

The test shall be completed using the randomizations providedby the experimenter. There shall be six randomizations for the sub-experiments, one for each group of four subjects. Each randomization shall be balanced across four blocks of 30 stimuli to eliminate long sequences of similar conditions or identical talkers. The sequences shall provide for alternating male-female talkers. Use of these randomizations will allow presentation order to be used as a factor in a global analysis, should that be necessary. The randomization shall be constrained to a randomized block design, which controls practice and fatigue effects that may occur over the course of a test session.

C10.8 Duration of the Experiment

Each trial consists of an eight-second reference sample + an eight-second test sample + five second voting time, totaling 21 seconds. For each of the four experiments there are 16 preliminary conditions x 21 seconds or 5.6 minutes for an introductory block. Each presentation set within an experiment consists of 60 conditions (A/B+B/A) x 4 talkers x 21 seconds or approximately 1h30min. The total testing time for each experiment will be 9 hours and 34 minutes, if four listeners are tested at one time.

Note that the above calculations do not include the time needed to give the subjects their instructions, or time taken for comfort breaks.

C10.9 Votes Per Condition

In each of the three experiments, 24 listeners rate every condition with four talkers in each of two presentation orders (A/B and B/A), giving:

(24 subjects x 4 talkers x 2 presentations) = 192 votes per condition

From past experience with CCR tests, this is the minimum number of votes per condition needed to give enough statistical certainty to differentiate the performance of one candidate process from another candidate process over the conditions and against the references.

C10.10 Test Procedure

Factors important for the experimental environment are specified in Sections 6.4, 6.5, and 6.6. Comfort breaks should be provided to reduce the effects of subject fatigue.

C10.11 Opinion Scale

The question asked of the subject is a based on of the CCR Listening Quality Comparison Scale. The listening subjects will judge the quality of the second sample with regard to quality of the first sample. The subjects will listen to each pair of samples and after these have been played, they will be asked to give their comparative opinion. Annex A contains an example of the instructions for the subjects in English. Changes to the instructions may be needed to specify the method of data collection being used (button-press, paper & pencil, etc.).

C10.12 Test Conditions for Experiment 4

Cond.	Noise	Input level	SNR (dB)	VAD/DTX	Reference	Processed	Speech sample
						Codec	number
1	Car	nominal	6	off	AMR@12.2	AMR@12.2	4 5 6 1 2 3
2	Street	nominal	9	off	AMR@12.2	AMR@12.2	4 5 6 1 2 3
3	Babble	nominal	9	off	AMR@12.2	AMR@12.2	4 5 6 1 2 3
4	Car	nominal	6	on	AMR@12.2	AMR@12.2	4 - - 1 - -
5	Car	nominal	6	N/A	Direct	MNRU-12	4 - - 1 - -
6	Car	nominal	6	N/A	MNRU-16	MNRU-12	4 - - 1 - -
4'	Street	nominal	9	on	AMR@12.2	AMR@12.2	- 5 - - 2 -
5'	Street	nominal	9	N/A	Direct	MNRU-12	- 5 - - 2 -
6'	Street	nominal	9	N/A	MNRU-16	MNRU-12	- 5 - - 2 -
4"	Babble	nominal	9	on	AMR@12.2	AMR@12.2	- - 6 - - 3
5"	Babble	nominal	9	N/A	Direct	MNRU-12	- - 6 - - 3
6"	Babble	nominal	9	N/A	MNRU-16	MNRU-12	- - 6 - - 3
7	Car	nominal	6	on	AMR@12.2	AMR/NS@12.2	1 2 3 4 5 6
8	Street	nominal	9	on	AMR@12.2	AMR/NS@12.2	2 3 4 5 6 1
9	Babble	nominal	9	on	AMR@12.2	AMR/NS@12.2	3 4 5 6 1 2
10	Car	low	6	off	AMR@12.2	AMR/NS@12.2	5 6 1 2 3 4
11	Street	low	9	off	AMR@12.2	AMR/NS@12.2	6 1 2 3 4 5
12	Babble	low	9	off	AMR@12.2	AMR/NS@12.2	1 2 3 4 5 6
13	Car	high	6	off	AMR@12.2	AMR/NS@12.2	2 3 4 5 6 1
14	Street	high	9	off	AMR@12.2	AMR/NS@12.2	3 4 5 6 1 2
15	Babble	high	9	off	AMR@12.2	AMR/NS@12.2	5 6 1 2 3 4
16	Car	nominal	15	on	AMR@12.2	AMR/NS@12.2	6 1 2 3 4 5
17	Street	nominal	18	on	AMR@12.2	AMR/NS@12.2	1 2 3 4 5 6
18	Babble	nominal	18	on	AMR@12.2	AMR/NS@12.2	2 3 4 5 6 1
19	Car	low	15	off	AMR@12.2	AMR/NS@12.2	3 4 5 6 1 2
20	Street	low	18	off	AMR@12.2	AMR/NS@12.2	5 6 1 2 3 4
21	Babble	low	18	off	AMR@12.2	AMR/NS@12.2	6 1 2 3 4 5
22	Car	high	15	off	AMR@12.2	AMR/NS@12.2	1 2 3 4 5 6

23	Street	high	18	off	AMR@12.2	AMR/NS@12.2	2 3 4 5 6 1
24	Babble	high	18	off	AMR@12.2	AMR/NS@12.2	3 4 5 6 1 2
25-48	Reversed order of the reference and processed speech samples in cond. 1-24						
Notes	- 4 talkers are used for all conditions: 2 male and 2 female - 6 speech samples (8 s) are used for each talker - 'multiple' conditions "4s", "5s" and "6s" (e.g. 4, 4' and 4") are only presented to a subset of listeners (e.g. to the first and the fourth groups of randomisation)						

C10.13 Statistical Analysis

The statistics to be reported from this CCR test are the averaged CMOS ($CMOS_k$) scores and the standard deviations (S_k) for all the conditions.

Additionally, the requirement in [1, Section 6.1.4] should be checked using hypothesis tests for the conditions 7-24 if the mean CMOS score is greater than zero (the NS performance is preferred) and greater or equal to zero (the NS performance is equivalent) within a 95 % confidence.

The hypothesis test should be performed using a 1-tailed T-test. The NS algorithm has failed the requirement at level "preferred" for any of test condition if

$$t < t_{N,0.05}$$

where

$$t = \frac{\mathrm{CMOS}_{test}}{\mathrm{S}_{test} / \sqrt{N}}$$

and the subscripts $_{test}$ denotes the test condition, N is the number of votes, and $t_{N,0.05}$ is the inverse of the Student's t-distribution with N degrees of freedom and probability 0.05.

Similarly, the NS algorithm has failed the requirement at level "equal" if

$$t < -t_{N,0.05}$$

C 11: Instructions to subjects and data collection

The instructions given to the subjects will to some extent depend on the method used to collect opinion data. In this section, example instructions are given for Pair Comparison, Modified ACR and CCR experiments. To ensure consistency, the actual instructions given to the subjects should be as close as possible to these examples, adapted for the number of speech files and length of the actual experiment, data collection method, and translated into the native language.

The instructions must be given prior to the commencement of the experiment, and the experimenter should ensure that the subject has understood them before starting the experiment. Questions asked by the subjects on procedural aspects of the experiment can, and should, be answered. However questions about the experiment design or what the experiment is investigating should not be answered until the subjects have completed the experiment. Subjects must be told not to give such information to subjects who are yet to participate in the experiment.

Subjects' responses may be collected by any convenient method: e.g. pencil and paper, press buttons controlling lamps recorded by the operator, or automatic data-logging equipment. Whichever method is used, care must be taken that subjects should not be able to observe other subjects' responses, nor should they be able to see the record of their own responses made in a previous session. Apart from the inevitable memory effects, each response should be independent of every other.

C11.1 Example Instructions for Experiment 1

In this test we are evaluating systems that might be used for a type of communications between separate places under a variety of conditions. You are going to hear a number of samples of speech reproduced in the earpieces of the handset. Each sample will consist of a sentence that was produced with two different communication systems. The first is identified as A and the second is identified as B.

Please listen to both A and B and then decide which of the two you prefer. Preference is strictly your decision and the decision should be based on your opinion of the quality of the speech samples. Some of the A/B pairs will seem clearly different and your decisions will be effortless. Others may be more difficult. ALWAYS MAKE A DECISION BETWEEN THE TWO. "I DON'T KNOW" and "I don't like either one" ARE NOT OPTIONS. Make your decisions independently. You should always compare the two current sentences, and not use any other presentation. Nor, should you be rating whether you like one talker better than another. This is not a test of you in any way; it is an evaluation of the systems. There is no right or wrong. Do not discuss how you are making your ratings during breaks or stretching periods.

For indicating your opinion you are requested to use the button box at your test station. <Use a prototype box to demonstrate with during training>. After listening to the two sentences, all lights on the box will flash. At that time, please press the appropriate single button that represents your opinion of the communication quality of the sample just heard. Use the leftmost button if you preferred the first sample (or A). Use the button on the far right if you preferred the second speech sample (or B). The corresponding light will be activated when a choice has been indicated. Once the button has been pushed, you will not be allowed to change your mind, so please respond carefully.

After you have given your opinion there will be a short pause before the next sample begins.

For practice, you will hear a series and provide an opinion on each; then there will be a break to make sure that everything is clear. An administrator will be in the room to answer questions. From then on you will have a break after each test block (approximately xx minutes). After the test block there will be a three -minute break during which you may leave the room. This series will continue for the duration of the test.

It is imperative that you do a good job with each rating by giving a true opinion of the communication system samples. The ratings you make later in the day are as important as those made earlier in the day. Please stay alert and do you best each time you make a decision.

Thank you for participating in this research. Feel free to ask any procedural questions at this time

C11.2 Example Modified ACR Instructions for Experiment 2

Instructions to the listeners

In this experiment we evaluate systems that might be used for telecommunication service between separate places.

You will hear speech samples reproduced in a telephone handset. Every sample consists of four short unconnected sentences in an environment with varying amounts of background noise. Your task is to indicate your opinion of the overall sound quality with respect to any unnatural sounds leading to unpleasant effects in the sample. Please make your judgement of the sample considering unnatural sound during the complete sample.

Use the following 5-point scale:

Excellent:	no unpleasant effects
Good:	slightly unpleasant effects
Fair:	somewhat unpleasant effects
Poor:	very unpleasant effects
Bad:	severely unpleasant effects

After each stimuli there will be a short pause for you to give your opinion. As a practice, you will first hear several samples and give an opinion on each. Then we will check that everything is clear before we start the test. Don't hesitate to ask questions if you have any. The experiment is divided in four parts with breaks in between. The parts last approximately 15 minutes each. Please do not discuss your opinions with the other participants in the experiment.

Thank you for your participation.

C11.3 Example Instructions for Experiment 3 and 4

INSTRUCTIONS TO SUBJECTS

In this experiment we are evaluating systems that might be used for telecommunication services.

You are going to hear through the handset pairs of speech samples, most of which have been recorded in different noisy environments (for example inside a car, in an office, or on the street). The first sample you will hear will be the reference sample. You will then hear the same sample again, but this time it will have passed through a telecommunications system. These samples will each consist of two short unconnected sentences.

You should listen carefully to each pair of samples. When they have finished, please record your opinion of the second sample with regard to the first one using the following scale:

Much better

Better

Slightly better

About the same

Slightly worse

Worse

Much worse

For practice, you will first hear [n] sample pairs and give an opinion on each. There will then be a short break to make sure that everything is clear.

From then on you will have a break approximately every [p] minutes. The test will last a total of approximately [q] minutes.

Please do not discuss your opinions with other listeners participating in the experiment.

C 12: Processing Tables

To be provided by the listening laboratory This shall be reported along with the results of the experiments.

C 13: Presentation Orders

To be provided by the listening laboratory This shall be reported along with the results of the experiments.

Annex D (informative): Change history

Change history							
Date	**TSG #**	**TSG Doc.**	**CR**	**Rev**	**Subject/Comment**	**Old**	**New**
03-2001	11	SP-010102	A001	4	Addition of test plan and tidying	8.0.0	8.1.0
03-2001	11	SP-010102	A002	1	Update of C code for objective measures for NS algorithm characterization	8.0.0	8.1.0
03-2001	11	SP-010102	A003	1	Correction of Annex A	8.0.0	8.1.0
					removal of empty table	8.1.0	8.1.1