Joint TSG-S4#10 - SMG11#15 Meeting
Helsinki, Finland
February 28th - March 3rd, 2000

S4/SMG11 Tdoc 143 /00

Draft TS 06.77 V1.10.0 (2000-021)

Technical Specification

Digital cellular telecommunications system (Phase 2+); Minimum Performance Requirements for Noise Suppresser Application to the AMR Speech Encoder (GSM 06.77 version 1.10.0)

European Telecommunications Standards Institute

Reference

Keywords

Global System for Mobile communications (GSM), speech, Adaptive Multi rate (AMR), Noise suppression

ETSI Secretariat

Postal address

F-06921 Sophia Antipolis Cedex - FRANCE

Office address

650 Route des Lucioles - Sophia Antipolis
Valbonne - FRANCE
Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16
Siret N8 348 623 562 00017 - NAF 742 C
Association ‡ but non lucratif enregistrÈe ‡ la
Sous-PrÈfecture de Grasse (06) N8 7803/88

X.400

c= fr; a=atlas; p=etsi; s=secretariat

Internet

secretariat@etsi.fr
http://www.etsi.fr

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETR†314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available **free of charge** from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://www.etsi.fr/ipr).
Pursuant to the ETSI Interim IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETR†314 (or the updates on http://www.etsi.fr/ipr) which are, or may be, or may become, essential to the present document.

Foreword

This ETSI Technical Specification (TS) has been produced by ETSI Special Mobile Group (SMG).
This specification specifies minimum performance requirements for Noise Suppression for the Adaptive Multi Rate (AMR) codec within the digital cellular telecommunications system.
The contents of this TS is subject to continuing work within SMG and may change following formal SMG approval. Should SMG modify the contents of this TS, it will be republished by ETSI with an identifying change of release date and an increase in version number as follows:

Version ?.x.y

where:

- y the third digit is incremented when editorial only changes have been incorporated in the specification;
- x the second digit is incremented for all other types of changes, i.e. technical enhancements, corrections, updates, etc.

1 Scope

This specification specifies minimum performance requirements for noise suppression algorithms intended for application in conjunction with the AMR speech encoder. 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 Normative references

References may be made to:

a) specific versions of publications (identified by date of publication, edition number, version number, etc.), in which case, subsequent revisions to the referenced document do not apply; or
b) all versions up to and including the identified version (identified by "up to and including" before the version identity); or
c) all versions subsequent to and including the identified version (identified by "onwards" following the version identity); or
d) publications without mention of a specific version, in which case the latest version applies.

A non-specific reference to an ETS shall also be taken to refer to later versions published as an EN with the same number.

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

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 this specification 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 is under study in SMG2 WPA].

5 Requirements to be assessed by Objective Means

5.1 Bit Exactness of the Speech Encoder

The Noise Suppression solution may be implemented as a pre-processing element situated ahead of and independent of the AMR speech encoder.

Alternatively the Noise Suppression algorithm may be implemented as an embedded module within the AMR speech encoder after the pre-processing module (sample down-scaling and high pass filtering) and operate on the pre-processed input speech buffer, denoted by "old_speech[L_TOTAL]" in the structure "cod_amrState" in the AMR C code [GSM 06.73: ANSI-C code for the GSM Adaptive Multi-Rate (AMR) speech codec].
The noise suppression algorithm is not allowed to modify any existing functions, tables, or internal variables of the AMR speech encoder except for the aforementioned speech buffer. (This is to be confirmed by SMG11.)

5.2 Use of AMR Speech Encoder functions

The Noise Suppression function shall have access to all AMR speech encoder variables and functions in the sense that it can use any variable or the output of any function of the speech encoder. This use of the AMR speech encoder variables and functions shall only be allowed on the condition that the speech encoder bit-exactness is preserved (as defined in Section 5.1).

5.3 Bit Exactness of the Speech Decoder

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

5.4 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. (This is to be confirmed by SMG11.)
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. (This is to be confirmed by SMG11.) 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.) .

5.5 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 ~~less~~ than the maximum of y and x. Any case where w is greater than y should be further investigated.
These requirements shall apply to all standardized AMR VADs. (w,x,y) are determined using all VADs, and 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 ([reference to be added when test plan is available]) .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 (*number of test*

conditions to be confirmed) conditions tested. Preference or equality within the 95 % confidence interval is required for the remaining conditions.
[Requirements for SNR improvement are for further study.]

Additionally, it is required that the subjective SNR improvement as measured by the methodology [Ref 1] where the measure is conducted on all CCR tests [Ref. 2] 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 beless than 4dB within the 95% confidence interval

[Note: Refs 1 and 2 to be added; Ref 1 references the SNR improvement measurement methodology, Ref. 2 references the test plan, currently under development, designed to test the requirements in this specification.]

7 Performance Objectives assessed by Objective Measures

7.1 Effect on Average Speech Level

[TBA]

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.

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~~2~~ dB, following otherwise the guidelines set forth in [Annex 1].	–7 dB or lower

Objective quality measure/test condition	Performance objective
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~~2~~ 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 1: 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 presented to be used for characterising NS solutions complying with the AMR/NS specification.

1 Objective Measures and Test Signals

1.1 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 $\mathbf{s_i}$, ***i = 1 to I***.
- The noise signals are referred to as $\mathbf{n_j}$, ***j = 1 to J***.
- The noisy speech test signals are referred to as $\mathbf{d_{ij} = \beta_{ij}(SNR)\, n_j + s_i}$, ***i = 1 to I, j = 1 to J***, where d_{ij} is built by adding s_i and n_j with a pre-specified SNR as presented below.
- The processed signal are referred to as $\mathbf{y_{ij} = NR\,(d_{ij})}$.
- 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 $\mathbf{c_{ij} = AMR\,(d_{ij})}$, ***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.
- $\mathbf{\beta_{ij}(SNR)}$ is the scaling factor to be applied to the background noise signal $\mathbf{n_j}$ in order to have a ratio **SNR** (in dB) between the clean speech signal $\mathbf{s_i}$ and $\mathbf{n_j}$. The scaling of the input speech and noise signals is to be carried according to the following procedure:
  - The clean speech material is scaled to a desired dBov level with the ITU-T recommendation P.56 speech voltmeter, one file at a time, each file including a sequence of one to four utterances from one speaker.
  - A silence period of 2 s is inserted in the beginning of each of the resulting files to make up augmented clean speech files.
  - Within each noise type and level, a noise sequence is selected for every speech utterance file, each with the same length as the corresponding speech files , and each noise sequence is stored in a separate file.
  - Each of the noise sequences is scaled to a dBov level leading to the SNR condition corresponding to the $\mathbf{\beta_{ij}(SNR)}$ value in each of the test cases by applying the RMS level based scaling according to the P.56 recommendation.
- The determination of which frames contain active speech is to be carried out with reference to the ITU-T recommendation P.56 active speech level measurement

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

1.2 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.

1.3 Proposal for objective measures for NS performance assessment

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

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

For each background noise condition j

For each speaker i

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

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

where β_{ij} depends on the SNR condition according to the procedure described in section 0

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

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

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

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

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

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

ξ is a constant that should be set at 10^{-5}

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

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

$SNRI\_n_{ij}$ correspondingly for frames at appr. the noise power level

$$SNRI_{ij} = \frac{1}{K_{sph} + K_{spm} + K_{spl}} \left(K_{sph} \cdot SNRI\_h_{ij} + K_{spm} SNRI\_m_{ij} + K_{spl} SNRI\_l_{ij} \right) \quad (2)$$

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

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

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:

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

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

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

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

for all signal frames k

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

$$\begin{aligned}
&\text{if } \mathrm{sp_pow}(k) \geq \mathrm{sp_lvl} + \mathrm{th_h} \\
&\quad \{k_{sph,\mathrm{length}(k_{sph})+1}\} = \{k_{sph,\mathrm{length}(k_{sph})}, k\} \\
&\text{else if } \mathrm{sp_pow}(k) \geq \mathrm{sp_lvl} + \mathrm{th_m} \\
&\quad \{k_{spm,\mathrm{length}(k_{spm})+1}\} = \{k_{spm,\mathrm{length}(k_{spm})}, k\} \\
&\text{else if } \mathrm{sp_pow}(k) \geq \mathrm{sp_lvl} + \mathrm{th_l} \\
&\quad \{k_{spl,\mathrm{length}(k_{spl})+1}\} = \{k_{spl,\mathrm{length}(k_{spl})}, k\} \\
&\text{else if } \mathrm{sp_lvl} + \mathrm{th_nl} \leq \mathrm{sp_pow}(k) < \mathrm{sp_lvl} + \mathrm{th_nh} \\
&\quad \{k_{nse,\mathrm{length}(k_{nse})+1}\} = \{k_{nse,\mathrm{length}(k_{nse})}, k\}
\end{aligned} \quad (9)$$

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.

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

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:

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

According to our experimentation, the results of the analysis are not highly sensitive to the selection of the threshold values. However, care has to be taken especially in the determination of the th_l and th_nh threshold values to avoid confusion between low power speech and a weak background noise present in the clean speech samples.

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

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

For each background noise condition j

For each speaker i

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

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

where β_{ij} depends on the SNR condition according to the procedure in section 0

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

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

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

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

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

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

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

~~**ETSI STC SMG11#11 Tdoc SMG11 408R/99**~~

~~Title: Description of revised new objective measures for assessing the SNR improvement and noise power level reduction produced by AMR/NS candidates~~

~~**Source: Nokia**~~

~~1 Scope~~

~~This document is a revision of a presentation of two new objective measures for assessing the performance of noise suppression (NS) methods. The presented measures have been accepted for being used to provide auxiliary information of the AMR/NS candidates in the selection phase. The results of the usage of the presented measures are part of the AMR/NS selection deliverables according to the selection deliverables document version 1.0 (SMG11 Tdoc 370/99). In the end of the document, some comments are recorded concerning the selection test source material as a test material for the proposed measures.~~

~~The version Tdoc 408R/99 incorporates two corrections: firstly the presentation of the NPLR was changed to be in accordance with the implementation in the objective measure tool provided by Nokia by exchanging the order in the subtraction of the noisy and noise reduced signals. Secondly, one inappropriate reference to an earlier used noise level speech class was removed.~~

~~2 NEW proposals for objective measures and TEST SIGNALS~~

~~2.1 Backgrund~~

~~The objective measures proposed in this document are based on an original proposal in SMG11 Tdoc 233/99, having a reference to Tdoc 153/98. The original proposal was refined in Tdocs 268/99 and 280/99, and some further notes have been made in the 8th AMR/NS sub-group meeting report (Tdoc 295/99) and afterwards over the SMG11-NS email reflector.~~

~~Some additional modifications have been made in both the objective measure calculation and the preferred conditions, which are included in the presentation in this document:~~

- ~~The notation of the total signal-to-noise ratio improvement measures have been changed from SNR_{imp} to ***SNRI*** and the corresponding measures for the high, medium and low power speech classes, or SIMP_h, SIMP_m and SIMP_l, to SNRI_h, SNRI_m and SNRI_l, respectively.~~

- A small change has been made in the calculation of each of the speech power class SNR improvement terms, as expressed in Eq. (1).
- The order of summation with regard to noise conditions and speech samples have been changed for both the ***SNRI*** and the ***NPLR*** measures.
- Division by frame length has been added in the speech frame power (sp_pow) calculation in Eq. (8) and the erroneaous scaling of the logarithm has been fixed from 20 to 10
- The threshold values for the classification of the signal frames into the power classes, Eq. (10), have been modified to correspond to the changed scaling in Eq. (8).
- The preferred test conditions have been further refined in this document in section 0.

2.2 Notations

The following notations are based on Tdoc 153/98, with some modifications.

- The clean speech signals will be referred as $\mathbf{s_i}$, ***i = 1 to I***.
- The noise signals will be referred as $\mathbf{n_j}$, ***j = 1 to J***.
- The noisy speech test signals will be referred as $\mathbf{d_{ij} = \beta_{ij}(SNR)\, n_j + s_i}$, ***i = 1 to I, j = 1 to J***, where d_{ij} is built by adding s_i and n_j with a pre-specified SNR as presented below.
- The processed signal will be referred as $\mathbf{y_{ij} = NR\,(d_{ij})}$, the operator NR(·) referring to the processing by the NS algorithm and the AMR speech codec.
- The reference signal in the calculations shall be either the noisy speech test signal d_{ij} itself or d_{ij} processed by the AMR speech codec without NS processing. The latter signal will be referred to as $\mathbf{c_{ij} = AMR\,(d_{ij})}$, ***i = 1 to I, j = 1 to J***, where the operator AMR(·) refers to processing by the AMR speech codec with no NS. The relevant reference signal will be indicated in the formulation of each objective measure below.
- The notation *Log*(·) indicates the decimal logarithm.
- $\mathbf{\beta_{ij}(SNR)}$ is the scaling factor to be applied to the background noise signal $\mathbf{n_i}$ in order to have a ratio **SNR** (in dB) between the clean speech signal $\mathbf{s_i}$ and $\mathbf{n_j}$. The scaling of the input speech and noise signals is to be carried according to the following procedure:
  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 $\mathbf{\beta_{ij}(SNR)}$ value in each of the test cases by applying the RMS level based scaling according to the P.56 recommendation.
- The determination of which frames contain active speech is to be carried out with reference to the ITU-T recommendation P.56 active speech level measurement and is related to the classification of the frames into the presented speech power classes which is explained below.
- 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.

2.3 Test material

The test material should manifest at least the following extent:

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

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

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

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

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

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

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

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

Note. There was a processing step in the validation study report, Tdoc 280R/99, noting that a 2 s initial convergence period was removed after processing from the test material. This step can, however, be omitted since the classification of the frames being based on the clean speech signal and on comparisons to the active speech level, no frames from the initial convergence period will be involved in any of the measurements.

2.4 Proposal for objective measures for NS performance assessment

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

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

For each background noise condition j

For each speaker i

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

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

where β_{ij} depends on the SNR condition according to the procedure described in section 0

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

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

$$\text{SNROut\_ij} = \frac{\frac{1}{K_{sph}} \sum_{k=k_{sph,1}}^{k_{sph,K_{sph}}} \sum_{n=k\cdot 80}^{k\cdot 80+79} y_{ij}^2(n)}{\frac{1}{K_{nse}} \sum_{l=k_{nse,1}}^{k_{nse,K_{nse}}} \sum_{n=l\cdot 80}^{l\cdot 80+79} y_{ij}^2(n)} - 1$$

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

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

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

SNRI_m_{ij} correspondingly for medium power frames

SNRI_l_{ij} correspondingly for low power frames

SNRI_n_{ij} correspondingly for frames at appr. the noise power level

$$\text{SNRI}_{ij} = \frac{1}{K_{sph} + K_{spm} + K_{spl}} \left(K_{sph} \cdot \text{SNRI\_h}_{ij} + K_{spm} \text{SNRI\_m}_{ij} + K_{spl} \text{SNRI\_l}_{ij} \right) \qquad (2)$$

$$\text{SNRI}_{j} = \frac{1}{I} \sum_{i=1}^{I} \text{SNRI}_{ij} \qquad (3)$$

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

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:

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

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

$$\text{SNRI\_l} = \frac{1}{J} \sum_{j=1}^{J} \text{SNRI\_l}_{j} = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{I} \sum_{i=1}^{I} \text{SNRI\_l}_{ij} \qquad (7)$$

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

for all signal frames k

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

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

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

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

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

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

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

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

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

where $\varepsilon > 0$ is a constant, such as 10^{-5};

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

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

- 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.
- We realise that in poor SNR conditions, the noise power level may occur to be higher than the lower bound of some of the speech power classes. However, even in this case, the information of the effect on the low power portions of speech may be informative. Naturally, another way of formulating the measure might be to make the power thresholds dependent on the noise level. This would, however, restrict the comparability of the SNR improvement figures of the different classes over experiments with different background noise content.
- The presented method of classifying the speech frames in the designated classes and, hence, determining values for the SNR improvement measures, is only applicable if all the used power level threshold values are higher than the corresponding power threshold level derived in the speech level measurement referred to above.

A preferrable scaling for the clean speech material is a normalisation to the active speech level of -26 dBov. In such a case, the following values are recommended for the power class thresholds:

th_h = -1 dB
th_m = -10 dB
th_l = -16 dB (10)

th_nh = -19 dB
th_nl = -34 dB

According to our experimentation, the results of the analysis are not highly sensitive to the selection of the threshold values. However, care has to be taken especially in the determination of the th_l and th_nh threshold values to avoid confusion between low power speech and a weak background noise present in the clean speech samples.

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

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

For each background noise condition j
For each speaker i
Construct a noisy input signal d_{ij} as follows:
$d_{ij}(n) = \beta_{ij} n_j(n) + s_i(n)$
where β_{ij} depends on the SNR condition according to the procedure in section 0
$c_{ij} = AMR(d_{ij})$
$y_{ij} = NR(d_{ij})$

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

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

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

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

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

3 Comments on the AMR/NS selection test material

In section 0, we have expressed the premise that the the street noise test material used in conjunction with the presented objective quality measures should be of a slowly varying power level. As a candidate proponent having gone through the processing of the source speech material with our AMR/NS candidate solution, we now have some experience on the noise material used for the AMR/NS selection tests. Our impression of the street noise material is not quite consistent with the requirement stated above, since the street noise samples appear to contain, to some

~~extent, background speech, horns and similar components whose frame power varies in a rate whose range coincides that of speech. Hence, the results to be obtained for the street noise conditions will have to be interpreted with special care.~~

~~4 On the scope of usage of objective measures for NS evaluation~~

~~Nokia want to point out that the objective measures presented in this document are intended for characterising some relevant aspects of the performance of NS algorithms and they might help in the comparison of AMR/NS candidates that are found equal by other means. However, we want to emphasise that the figures obtained with the proposed measures should be used as auxiliary information only. The subjective test results are to be used as the principal data for ranking the AMR/NS candidates in the selection process.~~

History

Document history		
V0.0.1	January 2000	First Draft