ETSI STC SMG11#15 **SMG11 Tdoc 112/00**

Helsinki, Finland

February 28th-March 3rd, 2000

Title: Objective measures for characterising the SNR improvement and noise power level reduction produced by NS algorithms

Source: Nokia

This is a proposed annex to the specification “Digital cellular telecommunications system (Phase 2+); Minimum Performance Requirements for Noise Suppresser Application to the AMR Speech Encoder (GSM xx.xx version 10.0.021)”

It presents an objective methodology proposed by Nokia for characterising the performance of noise suppression (NS) methods. Suggestions to the method obtained during the preparation of the AMR/NS selection phase in the AMR/NS subgroup of SMG11 have been taken into account in the formulation. Two objective measures are presented that were used to provide auxiliary information of the AMR/NS candidates in the selection phase and have been accepted by SMG11 to be used for characterising NS solutions complying with the AMR/NS specification. The rest of this document is intended to act as an annex in the AMR/NS minimum performance specification.

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 **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} = β_{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.
- **β_{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 **β_{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:
      $d_{ij}(n) = \beta_{ij}\ n_j(n) + s_i(n)$
              where $\beta_{ij}$ depends on the SNR condition according to the procedure
                described in section 1.1
    $c_{ij}$ = AMR ($d_{ij}$)
    $y_{ij}$ = NR ($d_{ij}$)
```

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

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

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

(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

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

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

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

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:

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

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

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

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

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

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

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

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

$$\left\{k_{spm,\mathrm{length}(\mathrm{k}_{\mathrm{sp}\,m})+1}\right\} = \left\{k_{spm,\mathrm{length}(\mathrm{k}_{\mathrm{sp}\,m})}, k\right\} \tag{9}$$

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

$$\left\{k_{spl,\mathrm{length}(\mathrm{k}_{\mathrm{sp}\,l})+1}\right\} = \left\{k_{spl,\mathrm{length}(\mathrm{k}_{\mathrm{sp}\,l})}, k\right\}$$

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

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

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:

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 $\beta_{ij}$ depends on the SNR condition according to the procedure in
              section 1.1
    $c_{ij}$ = AMR ($d_{ij}$)
    $y_{ij}$ = NR ($d_{ij}$)
```

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

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

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

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

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.