3QUEST: Comparison of EG 202 396-3 to TS 103 106

Contents

1. INTRODUCTION 4
2. DEVELOPMENT AND IMPLEMENTATION 5
2.1. STATUS OF EG 202 396-3 AND DISCUSSION IN 3GPP 5
2.2. NEW DATABASES 5
2.3. MODIFICATIONS OF THE ETSI EG MODEL 6
2.3.1. DATA TRANSFORMATION 6
2.3.2. PRE-PROCESSING 7
2.3.3. ALGORITHM CHANGES 8
2.4. PREDICTION RESULTS OF TRAINING DATA 10
2.4.1. NARROWBAND TRAINING 11
2.4.2. WIDEBAND TRAINING 12
2.5. VALIDATION PROCESS & PREDICTION RESULTS 13
2.6. NEW STANDARD: TS 103 106 15
3. ACQUA APPLICATION 16
3.1. MEASUREMENT WITH ACQUA 17
3.2. MANUAL ANALYSIS IN ACQUALYZER 19
4. COMPARISON TESTS 20
4.1. GENERAL REMARKS 20
4.2. NARROWBAND HANDSET TESTS 21
4.3. NARROWBAND HANDS-FREE TESTS 21
4.4. WIDEBAND HANDSET TESTS 22
4.5. WIDEBAND HANDS-FREE TESTS 23
4.6. STATISTICAL EVALUATION OF COMPARISON 23
4.7. CONCLUSIONS 26
5. SUMMARY 28
6. REFERENCES 29

© 2013 HEAD acoustics
Rev0 (06/2013)

1. Introduction

The objective model according to ETSI EG 202 396-3 [1] was developed to predict speech, noise and global quality of noisy speech signals for wide- and narrowband terminals according to ITU-T recommendation P.835 [2]. Especially for narrowband applications, signal processing capabilities of modern terminals (mobiles, smartphones) have rapidly progressed in the last years so that even 2-channel-microphone noise reduction solutions are currently state of the art. As a result, a much higher speech and noise quality can be achieved with these devices than without this technique.

Furthermore, it was discussed whether the underlying subjective narrowband data used in Annex H of EG 202 396-3 [1] from 2007 do not represent the latest technologies used for noise cancellation and it was suspected that there may be impairments potentially not covered correctly by the ETSI model.

In addition, some difficulties were reported [3] in achieving a reasonable correlation between a series of subjective tests and objective measurements following ETSI EG 202 396-3. It was found that this applies specifically to conditions which were not in scope of the ETSI databases, mainly conditions with different "SNR-sweeps" as described in Amendment 1 of ITU-T P.835 [2].

While revising the measurement standards TS 26.131 and TS 26.132 in 3GPP, it was agreed to include a method for objectively predicting speech, noise and global quality for handset terminals in sending direction and to possibly replace the ANR measurement by this new method. Due to the reported issues above, it was planned to create a modified version of the algorithm originally defined in ETSI EG 202 396-3 which should be adapted to a completely new set of subjective databases. After several parties contributed to this project with a huge effort of subjective testing, the model was retrained to this new material.

Due to the success of this process and the good prediction performance of the validation databases, the work of 3GPP was directly transferred into a new ETSI standard TS 103 106 [5].

This application note gives an overview about this new method and the consequences for the HEAD acoustics measurement system HEAD Analyzer ACQUA. Finally, several comparison measurements between the EG 202 396-3 and the TS 103 106 method are presented to provide an informative basis for the daily usage of both algorithms.

© 2013 HEAD acoustics
Rev0 (06/2013)

2. Development and Implementation

In this chapter, a brief description of the process of development and the algorithm modification is given. The following sections are intended to provide summary of the work which led to the new ETSI TS 103 106 standard, for further details refer to [5].

2.1. Status of EG 202 396-3 and Discussion in 3GPP

In the beginning of 2012, it was discussed to include a psycho-acoustically motivated test procedure for speech quality in background noise scenarios in the established measurement standards TS 26.131 and TS 26.132 for mobile phones. Prior to these discussions it became obvious that neither the existing ANR method nor other methodologies such as SNRI [7] were suitable to adequately predict the performance of advanced noise cancelling techniques used in modern mobile phones. It was evident that only a method which can predict the speech quality as perceived by the user could be a solution to this problem.

Since the only standard describing an objective method for this purpose was EG 202 396-3, the first proposal was to use this method for the next revision of the standard. But as already mentioned, some problems with the correlation of objective results of this method and subjective listening tests conducted by parties in 3GPP were reported [3]. The contexts of these listening tests always were state-of-the-art mobiles including multi-channel noise reduction systems.

As the need for an objective predictor of speech, noise and global quality found mutual consent, a new approach was chosen: several partners in 3GPP created new subjective listening tests and provided these databases to HEAD acoustics. The idea was to retrain the existing ETSI model to these new databases with as little changes as possible.

2.2. New Databases

The new subjective databases and the corresponding listening tests were created and conducted according to ITU-T recommendation P.835. This recommendationbest describes the basic procedure for conducting subjective tests and deriving the different quality dimensions S-MOS (speech mean opinion score), N-MOS (noise mean opinion score) and G-MOS (overall mean opinion score). Because of the rather general descriptions in this standard, a more detailed test plan was developed [6].

Several parties contributed with different types of databases which were divided into training and validation parts. Between 30 and 102 conditions per database were provided (including 12 reference conditions described in [6]), which led to an enormous amount of subjectively rated samples (8–24 samples per condition, 8–24 votes per sample).

© 2013 HEAD acoustics
Rev0 (06/2013)

An overview of the contributors and the extents of the databases is given in table 1. In particular, it is worth mentioning that well over 5000 single samples were made available for retraining in each mode (wide- and narrowband).

	Samples per condition	Narrowband				Wideband			
		Training		Validation		Training		Validation	
		(S)	(C)	(S)	(C)	(S)	(C)	(S)	(C)
Audience	8	1920	240	768	96	1440	180	768	96
Nokia						960	60	0	0
Orange France	12	0	0	0	0	1080	90	360	30
Qualcomm	16				96	1920	120	0	0
Total		5760	480	2304	192	5400	450	1128	126

Table 1: Amount of conditions (C) and samples (S) provided by several parties

2.3. **Modifications of the ETSI EG model**

A short overview of the changes of the EG 202 396-3 algorithm is presented in the following. For detailed descriptions of the modifications, please refer to [5].

2.4. **Data Transformation**

One of the most important differences between the new retrained model and the ETSI EG 202 396-3 method is the usage of multiple databases for the training. Due to the different spread of conditions amongst the quality scale, each listening test usually has its own relative quality dimensions because all the conditions are judged subjectively in the context of this individual listening test.

In order to better align different tests, a set of reference conditions as defined in [6] was used for these series of subjective tests so that the scales for S-, N- and G-MOS may eventually be set into a common context. This anchoring was applied with the reference conditions, which were available for each listening test database.

For each database, a mapping between the reference conditions and the average reference condition set is calculated. To catch also inter-relations between speech, noise and global ratings, a matrix transformation instead of a per-scale regression was chosen. To compensate for biases, a constant column was added to the reference set. Then, a transformation T_j is calculated for each database j with reference set R_j which minimizes the distance to the average reference set A:

© 2013 HEAD acoustics
Rev0 (06/2013)

$$\underbrace{\begin{pmatrix} 1 & S_{i01} & N_{i01} & G_{i12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S_{i12} & N_{i12} & G_{i12} \end{pmatrix}}_{R_j(\text{Ref. set } j)} \times T_j = \underbrace{\begin{pmatrix} \overline{S_{i01}} & \overline{N_{i01}} & \overline{G_{i12}} \\ \vdots & \vdots & \vdots \\ \overline{S_{i12}} & \overline{N_{i12}} & \overline{G_{i12}} \end{pmatrix}}_{A\ (\text{Avg. ref. set})}$$

The transformation matrix T_j (size 4 × 3) can easily be determined to be

$$T_j = \left({R_j}^T \times R_j\right)^{-1} \times {R_j}^T \times A.$$

If the three scales (S-MOS / N-MOS / G-MOS) are independent from each other for each database, the matrix transformation T_j equals a linear per-scale transformation. Before the retraining of the model, the transformation is applied to the whole test data on a per-sample base:

$$\underbrace{\begin{pmatrix} 1 & S_1 & N_1 & G_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S_N & N_N & G_N \end{pmatrix}}_{S_j\ (\text{scores of samples of database } j)} \times T_j = \underbrace{\begin{pmatrix} \tilde{S}_1 & \tilde{N}_1 & \tilde{G}_1 \\ \vdots & \vdots & \vdots \\ \tilde{S}_N & \tilde{N}_N & \tilde{G}_N \end{pmatrix}}_{\tilde{S}_j\ (\text{transformed scores of samples of database } j)}$$

After transforming all scores individually per database, they could then be used for the retraining of the ETSI model. In consequence, all MOS values predicted by the new algorithm are also derived from this "average" context which is a mix of several databases.

2.5. Pre-Processing

Pre-Filtering (NB)

In the narrowband extension of [1], the listening test audio files included a far-end handset simulation, realized with an *intermediate reference system* (IRS) RCV filter according to annex D of [8]. In the requirements described in [6], such a listening filter was described or used in the databases neither for narrow- nor for wideband.

The narrowband mode of ETSI EG 202 396-3 internally filters the unprocessed and clean reference with IRS SND and IRS RCV to simulate a transmission over high-quality listening devices and networks. The principle of IRS seems to be outdated, modern state-of-the-art mobiles do not have this frequency characteristic. In these newly created NB databases, the used devices show almost flat frequency responses in sending direction.

Thus, the filtering with IRS SND and RCV of the two reference signals was replaced by filtering with the *mobile station in* (MSIN) filter [10], which is mainly a band pass. A listening filter is not applied to the processed signals in the retrained model.

© 2013 HEAD acoustics
Rev0 (06/2013)

Speech part detection (NB & WB)

The detection of signal parts belonging to either speech or noise was updated. Now the clean speech signal is segmented into frames and classified according to G.160 [7]. The signal parts classified as silence are assumed as background noise sections, all other frames are assumed as speech.

Speech level adjustment (WB)

According to [6], the level adjustment of the recordings of the training databases was applied in such a way that the active speech level over the full sequence test must be set to 73 dB SPL (-21 dB Pa) for the listening test.

The EG 202 396-3 implementation assumes 79 dB SPL (-15 dB Pa) active speech level due to the underlying listening test. Thus, this constant was adapted to the new databases.

2.6. Algorithm Changes

The ETSI model calculates several parameters out of the psycho-acoustically motivated inner representation for the estimation of S- and N-MOS. The parameters for S-MOS are presented in table 2. A detailed description of the calculation for the parameters can be found in [1].

The calculation of the objective S-MOS in chapter 6.5.2 of [1] is performed with a linear quadratic regression of the parameters mentioned above. In addition, the regression coefficients are switched with regard to the N-MOS calculated before which models the listeners' expectation to speech quality.

$P_1 = \Delta\mathrm{SNR}$	$P_4 = \mu(\Delta RA_{\mathrm{Sp,P-C}})$
$P_2 = \mu(RA_{\mathrm{Sp,P}})$	$P_5 = \sigma^2(\Delta RA_{\mathrm{Sp,P-C}})$
$P_3 = \mu(\Delta RA_{\mathrm{Sp,P-U}})$	$P_6 = \sigma^2(\Delta RA_{\mathrm{Sp,P-U}})$

Table 2: Extracted Parameters for S-MOS

The applied modification is the replacement of the linear quadratic regression with a feed forward neural network. In consequence, the switching of the regression coefficients depending on the N-MOS is removed. Only one network is trained with input (six parameters of table 2) and output (S-MOS) data by a simple back-propagation algorithm [9].

© 2013 HEAD acoustics
Rev0 (06/2013)

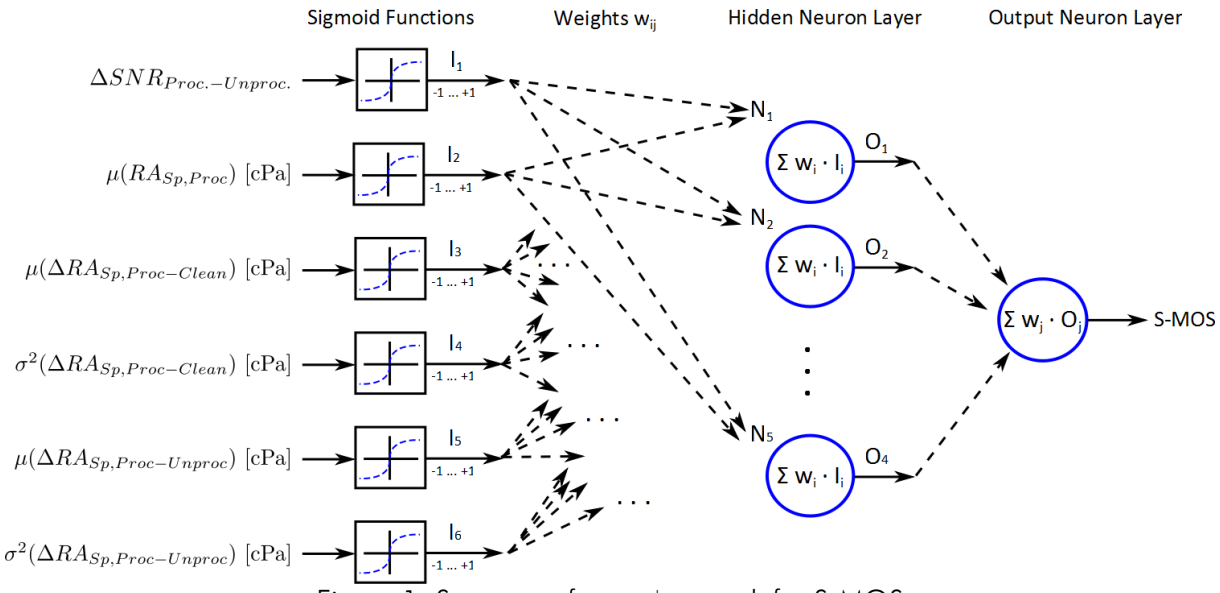


Figure 1: Structure of neural network for S-MOS

The setup of the neural network is shown in figure 1. It consists of 5 hidden layers; each layer N_j includes a connection from each transformed input parameter I_i. The output O_j of each layer is calculated as the weighted sum of each input I_i using the weights w_{ij}. The outputs O_j are then weighted by w_j and summed up to the output S-MOS. Both, w_{ij} and w_j are the result of the training of the network. The parameters according to table 2 are composed to a vector $\mathbf{P}$ including a bias as the first element:

$$\mathbf{P} = (1 \quad P_1 \quad P_2 \quad P_3 \quad P_4 \quad P_5 \quad P_6)$$

The output calculation of the neural network shown in figure 1 can be described as concatenated matrix operations:

$$S\text{-}MOS_{\text{objective,raw}} = f_{\text{sigmoid}}\left(f_{\text{sigmoid}}\left(\frac{\mathbf{P}-\mathbf{M}_{\text{in}}}{\mathbf{S}_{\text{in}}}\right) \times \mathbf{H}\right) \times \mathbf{O}$$

First the parameter vector $\mathbf{P}$ is normalized to mean 0.0 and standard deviation 1.0. This is done by subtracting the average of all training data for each parameter from each item of the input parameter vector. The averages for each parameter P_i can be described as a vector, which is different for narrow- and wideband mode. The numeric content of these vectors described can be found in [5].

After normalizing the input data, the sigmoid function $f_{\text{sigmoid}}(x)$ is applied to each normalized parameter P_i. This ensures that each input of each neuron of the hidden layer is soft-limited to the range ± 1.0 and guarantees that parameters out of the training range cannot produce an overflow which results in eventually unreasonable scores. For the current model, the hyperbolic tangent was chosen to a sigmoid function:

$$f_{\text{sigmoid}}(x) = \tanh(x)$$

Thus the input of the hidden neuron layers can also be given as a transformed vector $\widetilde{\mathbf{P}}$ of parameters:

© 2013 HEAD acoustics
Rev0 (06/2013)

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/268133140135006052