INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2017-2020

SG12-C.384

STUDY GROUP 12

Original: English

Question(s): Q9/12 Geneva, 7 – 16 May 2019

CONTRIBUTION

Source: TU Berlin

Title: Proposal for the Requirement Specification of P.SAMD

Purpose: Proposal

Contact: Gabriel Mittag
Quality and Usability Lab, TU Berlin
Germany
Tel: +49 30 8353 54246
Fax: +49 30 8353 58409
Email: gabriel.mittag@tu-berlin.de

Contact: Sebastian Möller
Quality and Usability Lab, TU Berlin
Germany
Tel: +49 30 8353 584645
Fax: +49 30 8353 58409
Email: sebastian.moeller@tu-berlin.de

Keywords: P.SAMD; single-ended; speech quality

Abstract: This contribution contains a proposal for the requirement specification of the single-ended and diagnostic speech quality prediction model P.SAMD. Besides the prediction of the four perceptual dimensions of P.AMD set A, the specification also includes the prediction of overall speech quality, as assessed in a P.800 absolute category rating listening-only test. We further propose to change the bandwidth mode of the model from super-wideband to fullband. Since P.SAMD is a single-ended model, without access to the clean reference, the training and testing databases for this work item should be based on conversational speech samples.

At the last SG12 meeting in November 2018, we presented first results of the P.SAMD model for the prediction of overall speech quality. Since the current single-ended speech quality model P.563 is only available for narrowband scenarios, we proposed to include the overall quality to the work item P.SAMD.

In this contribution, we propose a draft text for the requirement specification of P.SAMD that is based on the specification text of the multi-dimensional, reference-based work item P.AMD and the discontinued single-ended overall speech quality work item P.SPELQ.

To be in line with the current reference-based recommendation for overall speech quality assessment P.863, we propose to change the bandwidth of the P.SAMD model from super-wideband to fullband, with a maximum MOS score of 4.8 for a clean fullband speech signal.

Major modifications to the specification of P.AMD or P.SPELQ that may be discussed during the meeting are highlighted in the text.

Content

1 Scope of the objective model P.SAMD

The purpose of the model P.SAMD is to predict the overall speech quality, in narrowband, wide-band, super-wideband, and fullband telecommunication scenarios as it would be scored in a P.800 Absolute Category Rating (ACR) Listening Only Test (LOT) in a fullband context. In contrast to P.862 and P.863, the approach of P.SAMD is “single-ended” or “non-intrusive”, which means that the quality prediction is based on the received speech signal only. In addition to the single-ended speech quality prediction, the model also provides more detailed information about the cause of quality degradation with an approach based on perceptual quality dimensions. The quality dimensions are assessed in a listening-only test, following the procedure outlined in Section 2.4.

The evaluation procedure will be based on the overall quality, as well as each single perceptual dimension score by a statistical evaluation procedure. The evaluation terms are described in chapter 4.

Fullband operational mode

P.SAMD should allow only one operational mode. In this mode the speech samples are scored against a fullband (50 to 20 000Hz) reference signal and predict the perceptual dimension scores on a corresponding scale.

Restrictions

P.SAMD is not intended to score a per-call quality or quality for longer sequences of speech. It is focused on prediction of quality of shorter speech utterances such as 6 … 12s in length.

Other dimensions of speech quality such as conversational aspects and talking quality are not within the scope of P.SAMD. The P.SAMD model should consider noises and their influence on perceptual quality dimensions in a listening-only context similar to the one described in P.800 (test cabinet specifications, etc.). The prediction of quality as it can be perceived in a noisy listening environment and the related binaural effects are not in the scope of P.SAMD.

P.SAMD predicts the listening quality of human speech, including in the presence of background noises; however, it is not intended to rate the quality of synthetic speech, music, noise or other non-speech signals (e.g. signalling tones) alone. The detection and classification of such pure non-speech signals has to be done up-front to the P.SAMD model and will not be processed further.

Lip, breath and similar noises as well as filling sounds produced by the talker are considered as belonging to the speech signal.

Intended Scenario

Compared to common full-reference approaches, where a dedicated test connection has to be established and a test speech signal is transmitted, a no-reference approach can be applied to unknown voice signals. Since a no-reference approach does not require a dedicated test connection, it is able to predict listening quality of real live calls in monitoring scenarios.

Here two main applications are present; those are (a) predicting the quality of voice signals from a source such as e.g. a voice service like a weather forecast or (b) predicting listening quality in live conversations.

For both it has to be stated that the source speech signal cannot be pre-assumed as high quality and clean, potential degradations and environmental noises in the source will be part of the prediction.

The term *'telecommunication scenario'* as mentioned above covers all transmission technologies in today’s

- Public switched networks (e.g. fixed wire PSTN, GSM, WCDMA, CDMA, …),
- Push-over-Cellular, Voice over IP and PSTN-to-VoIP interconnections, Tetra and

- Commonly used speech processing components (e.g. codecs, noise reduction systems, adaptive gain control, comfort noise and other types of voice enhancement devices) and their combinations.

Other technologies or components such as speech storage formats or non-telephony applications such as public safety networks or professional mobile radio connections are not part of the competition and the selection criteria.

1.1 Perceptual dimension results provided by P.SAMD

Besides overall quality, the P.SAMD model will deliver separate quality scores for different types of distortions. Four dimensions have been identified which allowed diagnostic information to be obtained from speech signals, with a differing level of detail:

1) Coloration (i.e. resulting from frequency response distortions, e.g. bandwidth restrictions and coloration introduced by transducers)
2) Noisiness (e.g. resulting from additive and multiplicative noise)
3) Discontinuity (e.g. resulting from time localized and time-varying degradations)
4) Loudness (e.g. impact of overall play back level)

The evaluation will be done for each dimension individually against pre-defined objectives.

1.2 Test and applications scenarios for and P.SAMD

The so-called *'test scenario'* describes a set of individual connections or complete processing chains, which are typical for types of telephony applications (e.g. wireless connections in CDMA, simulated codec transmissions, hands-free terminals).

The *'application scenario'* means in this context the anticipated usage of the P.SAMD models that has already been considered in the design of the databases and the corresponding auditory tests. It covers different types of measuring interfaces or terminal types.

The *test scenarios* should cover commonly used so-called 'simulated' connections from e.g. codec standardisations as well as an equivalent amount of data recorded in real field scenarios. Both sets of data collections should reflect the actual behaviour in today's networks and telephony services. Thus, the latest coding technologies have been taken into account as well as typical telephony services such as video-telephony in 3G networks.

The *test scenarios* should be based on clean (noise-free) speech transmissions as well as the transmission of noisy speech (or the insertion of noise into the system under test). The way that noisy speech is handled by codecs or noise reductions systems is especially interesting and the influence of the residual noise on the perceptual dimension scores should be tested.[1]

The *test scenarios* should cover the following distortion types:

- Single and tandemned speech codecs as used in telecommunication scenarios today
- Packet loss and concealment strategies (packet switched connections)
- Frame- and bit-errors (wireless connections)
- Interruptions (such as un-concealed packet loss or handover in GSM)
- Front-End-clipping (temporal clipping)
- Amplitude clipping (overload, saturation)

---

[1] It might be necessary to force the operation of some of the functionality of the speech processing components by transmitting noise in the opposite direction of the connection.

- Effects of speech-coding systems on pre-noised speech
- Variable delay (VoIP, video-telephony) / Time Warping
- Gain variations
- Influence of linear distortions (spectral shaping), also time variant
- Voice enhancement systems in networks and terminals and their effects on Listening Quality

1.3 Capturing interfaces

P.SAMD requires an electrically recorded signal at either an endpoint interface or at a terminal device. Acoustically captured speech signals are not in the scope of P.SAMD.

As an end-point application a measurement point can be imagined as a network termination point (NTP) where the user's device is connected to or could be. This can be e.g. an ISDN interface but also a headset connector of a mobile handset.

In the end-point application, for example a mobile phone, P.SAMD may be used to evaluate the quality of received speech at the end-user's device, which becomes very close to the subjective perception of the listener.

Simulations and dedicated test connections

- Simulations: Pre-recorded source speech signals (which may include background noise) inserted into a simulated channel or other chain of speech processing (as e.g. for codec standardization).

- Test connections (electrical insertion): Pre-recorded source speech signals (which may include background noise) inserted into a live channel electrically with or without using a real sending terminal or soft-phone client.

- Test connections (live talking): A human talker is speaking into a real terminal's microphone in a quiet or noisy environment in combination with a live channel. This configuration is the most important regarding the scope of P.SAMD.

1.4 Development procedure and requirements on participants

The P.SAMD development and standardization process is a collaborative approach, where interested parties contribute to a joint development.

The collaborating parties will inform Q9/12 on a regular basis, at least at each ITU-T SG12 meeting, about the progress of P.SAMD. Potential changes in the time schedule and deadlines have to be agreed within Q9.

Each participant accepts the defined requirements for P.SAMD and the design rules for databases described in this document

- Each participant shall contribute speech databases to the collaboration as defined under 'Databases and Test plans' in this document. The workload should be shared equally between the participants. The legal discussion about the exchange of data bases is outside of ITU-T and to be agreed in multi-lateral cross-license agreement.
- Each participant has to be a member of the ITU-T SG12 or must be represented by a member of ITU-T SG12.

- Patent pooling, IPR, Know-How sharing and cross-licensing terms have to be discussed between participants outside of ITU-T.

As model ITU-T considers a system that produces the required output values based on the defined inputs.

Time schedule for the P.SAMD standardization process

1. Opening Work Item P.SAMD Sept, 2016
2. Revision of Requirement Specification to P.SAMD May, 2019
3. First version of P.SAMD model and result analysis Nov, 2019
4. Final model, evaluation results, Draft Recommendation P.SAMD Mid 2020

1.5 Characterization phase

After the development is finalized, it is agreed to characterize the model further by additional data in a separate characterization phase. It may underline strength of the upcoming Recommendation but also discover weaknesses of it. This characterization phase is good practice for codec standardization processes and will lead to well tested and properly described Recommendation.

It is intended that this characterization phase is conducted by the other proponents and independent third parties voluntarily. The characterization phase may include databases especially designed and agreed for this characterization purpose but in addition data bases available for the involved parties only. All results obtained and intended to be considered in this characterization phase have to be published within ITU-T SG12 for discussion. Detected and published weaknesses of the model by use of secret and unpublished databases have to be confirmed by an independent party in prior of consideration in the Recommendation. Based on the available results the scope and application scenarios of the upcoming Recommendation can be re-defined. In case of mal-function in certain conditions a bug-fix procedure can be agreed by ITU-T SG12.

This characterization phase should not exceed four months after the development is finished.

If no parties submit results within the characterization phase, the results of the development will be used for characterization.

2 Technical requirements and use cases of P.SAMD

This section describes the technical use cases for P.SAMD. The schemes and restrictions are valid for generating test data for the evaluation process as well as in real-live use of the selected P.SAMD model.

2.1 Fullband scenarios

Fullband scenarios

A requirement for P.SAMD is the assessment of fullband (50…20000Hz) in the corresponding applications. Furthermore, intermediate band-limitations on signals have to be scored adequately in a fullband scenario. Different presentation levels in the listening test have to be considered as test cases.

2.1.1 Technical Requirements on signals to be processed by P.SAMD

Input:

- Signals:
  - Speech signal format either WAV or RAW
  - Mono, 16bit linear PCM, Intel
  - Sampling frequencies: 48kHz
  - Length 6 to 12s
  - RAW format files must not use the file extension WAV

- Control parameters:
  - Sampling frequency in Hz (in case of WAV this value is ignored)

Technical requirements (for evaluation):

- Executable running under Windows 10 OS
- Command line operation

2.1.2 Predicted scores by the model

- Overall listening quality and perceptual quality dimensions Noisiness, Coloration, Discontinuity, and Loudness on a 1 to 5 MOS scale, with upper boundary 4.8
- Operational mode: fullband

2.1.3 Requirements for the model output parameters

It is proposed that the P.SAMD candidate has a tab separated output structure. Each individual evaluation should produce one single line. The results line should present the following results in the order:

degraded file name (or path)

Predicted overall quality

Predicted perceptual dimension value for 'Coloration

Predicted perceptual dimension value for 'Noisiness'

Predicted perceptual dimension value for 'Continuity'

Predicted perceptual dimension value for 'Loudness'

2.1.4 Scoring of Background noises

It is agreed that noise at the sending side or inserted in the transmission chain will be a test condition for P.SAMD.

The prediction of perceptual quality dimensions as it can be perceived in a noisy listening environment and the related binaural effects are not in the scope of P.SAMD.

2.4 Subjective listening-only test procedure for the four dimensions

The perceptual dimension scores predicted by P.SAMD should reflect as closely as possible the judgments of humans with respect to the individual dimensions given in Section 1.1. These judgments should be collected in a test carried out in the way described hereafter.

Because P.SAMD also predicts the overall quality according to a P.800 ACR listening-only test, the overall quality on an ACR scale may be rated by the test participants prior to the rating of the perceptual quality dimensions.

The test procedure for the dimensions is similar to a standard ACR overall quality test as it is described in ITU-T Rec. P.800 and P.830. All deviations from the procedure described in these two Recommendations are given hereafter.

Four descriptive scales are used for measuring the three dimensions “discontinuity”, “noisiness”, “coloration”, and “loudness”. That way, separate scores for the perceptual dimensions present in test conditions containing multi-dimensional degradations can be obtained. The graphical layout of the “discontinuity”, “noisiness”, “coloration” and “loudness” scales is similar to the scales recommended in ITU-T Rec. P.851 (2003). The poles of the scales are labeled with the antonym attributes “continuous – discontinuous” (discontinuity dimension), “not noisy – noisy” (noisiness dimension), “uncolored – colored” (coloration dimension), and “optimum level – non-optimum level” (loudness), see Figures 3-6.

Figure 3: "Noisiness" scale, German version. English translations: “not noisy”, “noisy”.

Figure 4: "Discontinuity" scale, German version. English translations: “continuous”, “discontinuous”.

Figure 5: "Coloration" scale, German version. English translations: “uncolored”, “colored”.

Figure 6: "Loudness" scale, German version. English translations: “optimum loudness level”, “non-optimum loudness level”.

In the dimension assessment experiment, the scales are presented separately, i.e. consecutively for each stimulus. In prior to the registering of the ratings, the listeners are asked to listen to the entire speech sample. During one trial, they can optionally repeat the playback. The rating scheme for one sample is depicted in Figure 7 (for three scales).

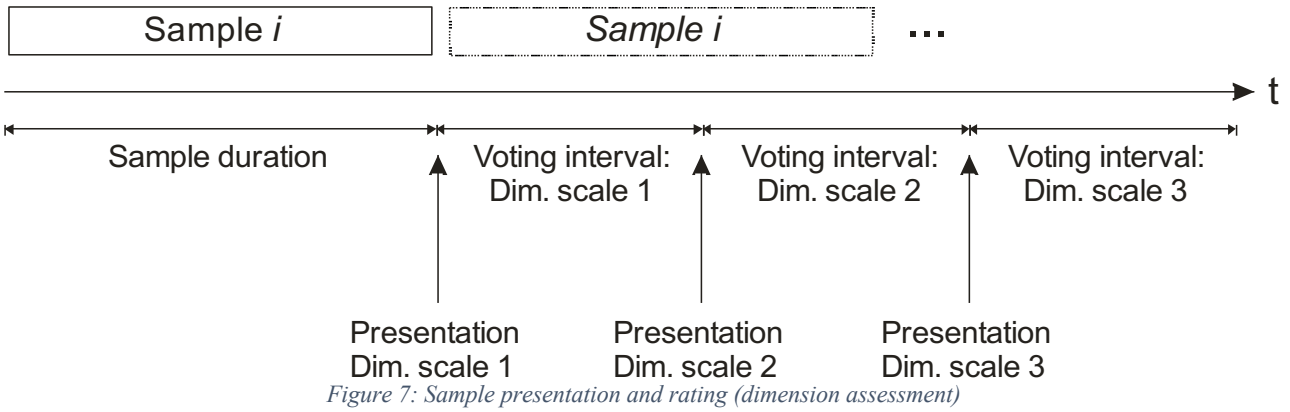

Figure 7: Sample presentation and rating (dimension assessment)

The samples are presented in randomized order. For each participant, the order of the scales is permutated, following the scheme tabulated in Table 1.

Table 1: Presentation order of the discontinuity, noisiness, coloration and loudness scale

Participant	Dim scale 1	Dim. scale 2	Dim. scale 3	Dim. scale 4
1	dis	noi	col	lou
2	noi	col	lou	dis
3	col	lou	dis	noi
4	lou	dis	noi	col
…	…	…	…	

The order is held constant per participant in order to avoid confusion of the scales.

A detailed description of the four dimension-scales is given to the subjects. The instructions start off explaining that in this part of the experiment, the *features* or *characteristics* of speech samples are supposed to be judged (i.e., not the quality), and that this evaluation is done by means of four scales. Each scale is labeled with an attribute at each end that describes the characteristic to be judged upon. Each scale and its usage are separately described in detail, using synonyms to the scale attributes as an aid. In detail, the participants are instructed that

- with the scale in Figure 3, the "noisiness" of the sample is supposed to be judged; the labels "not noisy" and "noisy" can be paraphrased with the terms "not hissing" and "hissing", respectively
- with the scale in Figure 4, the "discontinuity" of the sample is supposed to be judged; the labels "continuous" and "discontinuous" can be paraphrased with the terms "regular" / "steady" / "not chopped" / "not bubbling" / "not ragged" and "irregular" / "shaky" / "chopped" / "bubbling" / "ragged", respectively
- with the scale in Figure 5, the "coloration" of the sample is supposed to be judged; the label "uncolored" and "colored" can be paraphrased with the terms "direct" / "close" / "thick" / "not nasal" and "indirect" / "distant" / "thin" / "nasal", respectively
- with the scale in Figure 6, the "loudness level" of the sample is supposed to be judged; the label "optimum loudness level" means that the loudness level is neither too high nor too low.

The complete instructions can be found in COM12-C82, Appendix II, extended by the inclusion of the "loudness" scale, see the last bullet point.

The training for the dimension assessment is divided into

- a training phase where the listeners can internalize the *meaning* of the scales by acoustic examples, and
- a training phase where the listeners can familiarize with the *usage* of the scales.

For the training of the *meaning* of the scales, exemplary samples for each scale are presented which are distorted in (mainly) one dimension. Therefore, unidimensional "anchor conditions" specified in Section 3.4.1 corresponding to the four dimensions can serve as training samples.

The acoustic presentation is done together with the describing synonyms by means of a computer screen where the participants can listen to the samples until they confirmed that they understood the meaning of the scales. The understanding is supported by presenting an undistorted sample (direct SWB), stating that this particular sample is completely "not noisy", "continuous", "uncolored" and of "optimal loudness". A screenshot of the graphical training interface is included in the instructions in COM12-C82, Appendix II, and the software can be provided to interested laboratories upon request.

Preceding trials help the listeners familiarize with the practical *usage* of the scales and the range of degradations to be expected. Therefore, several samples differing in quality and character of the degradation were rated in a brief dedicated training session.

Further details regarding the test duration, the detailed organization of the experiment, and the complete instructions can be found in COM12-C82.

It has to be agreed upon a common transformation rule of the raw scores to "dimensional MOS-LQS" values. The raw ratings of the "discontinuity", "noisiness", "coloration" and "loudness" scores (ranging from 0 to 6 according to the scale design) can either be linearly transformed to the MOS-range [1;5], or the transformation rule according to the equation in Section 4.2.1 in COM12-C82 can be applied.

The resulting "dimensional MOS-LQS" values are denoted as MOS-LQS-dis, MOS-LQS-noi, MOS-LQS-col, and MOS-LQS-lou.

3 Databases and test-plans

There should be at least 5 training and 3 test databases of approx. 60 conditions available.

This would result in 8 databases, including around 8*60=480 conditions. The distribution of dimensions within these conditions (i.e. if a condition mainly reflects noisiness, loudness, or a combination of dimensions) should be further discussed, but it should be ensured that each quality dimension is properly exercised in the databases.

The POLQA and P.AMD cross-license databases are available also for the P.SAMD standardization efforts. This would provide a number of corresponding speech files, however yet without the subjective labels for the perceptual dimensions. It was further agreed that this pool is also used for exchange of P.SAMD databases.

Besides the P.SAMD databases that contain conversational reference speech files, focused on single-ended models, also the databases from the P.OLQA and P.AMD pool, with common P.800 sentence pairs can be used for training and testing. However, at least two of the three test databases should be databases with conversational reference files.

To train the model, experiments that are carried out in the crowd according to P.808 may be used.

It would be appreciated if third parties are contributing databases (or certain conditions of a database only) to the P.SAMD development process. This can be done either on a multi-lateral agreement or by using the Q9 speech pool.

3.1 Structure of new conducted full-scale tests in a fullband context

The new full-scale fullband databases have to follow the defined design rules and to be compliant with the minimum requirements in this section.

All those experiments can be considered as 'full-scale databases', which means that they must cover the entire range of degradation dimensions and background noise as well as clean conditions.

- Each file must be assessed by at least 8 subjects
- Each condition must include at least four talkers
- Each condition must be assessed by 96 votes, but not less than 72 and not more than 160.

3.1.1 Anchor conditions

The anchor conditions for P.SAMD databases are as follows:

- FB clean
- FB P50MNRU 25dB
- FB White Noise 12dB
- FB Level -20dB
- FB BP 500-2500Hz
- FB BP 100-5000Hz + Level -10dB
- FB Time Clipping 2%
- FB Time Clipping 20%

These conditions showed to have a good distribution of MOS values for all four dimensions in previous experiments.

3.1.2 Design rules of test conditions for full-scale tests

Each of the 'Full Scale' experiments must include:

- Background Noise (>30% but not more than 70% noisy), see definition of Background Noise in section 3.4.4
- Audio band-limitations (>10% SWB, >20% WB, >20% narrowband)
- Presentation level
    - >10% -20 … -12dB
    - >10% -12 … -3dB
    - >60% -3 … +3dB
    - >10% +3 …+6dB

The limitations given above have to be applied to the test conditions only (except anchor conditions).

Further degradation types are required for the new full-scale databases. Here the objectives should be reached on average over all full-scale databases used. This allows individual databases to focus on certain degradation types and to not include some types at all. The objectives have to be met without considering the anchor conditions.

- At least 40% live and at least 40% simulated network situations
- At least 15% Variable delay (VoIP, video-telephony) / Time Warping
- About 5% Amplitude clipping (overload, saturation)

- At least 60% Speech codecs as used in telecommunication scenarios today and cascading of those. Codecs are split over four different classes e.g. transform, CELP, waveform, sinusoidal
- 5 different PLC strategies; at least 15%
- 5 different packet loss patterns, including front-end clipping (temporal clipping); at least 15%

3.1.3 Live Talking

In live talking, a realistic conversation is recorded in real-time at far end and optionally at near-end, and there is no possibility to control the "source" of the speech utterance and the "transmission" of the speech utterance, but it is much more related to realistic scenario compared to above recording scenario.

The facilities (as a sender and receiver) could be smartphones, telephones, VoIP terminals with human talkers. When a (conversational) connection is created, the smartphone or another terminal (as a receiver) can get the real-time speech data (downlink) and record and store it. It can be done on the UE itself or by other means as e.g. recording the speech signal from an electrical interface in real-time during the conversation. Additionally, the probe can also be installed at one side to record only one direction.

In general, the collected data from above procedure are long-term speech (e.g., a speech file with several minutes duration). Therefore, we need to extract some short segments from the collected data manually.

To cover the realistic degradation condition and avoid infringing privacy, the requirements on this scenario include that ALL test persons should be informed in advance. Additionally, all contents in conversations don't include improper words and contents as required for public listening tests.

It is recommended to initiate the conversation by established scenarios for conversational tests as described in P.805 (structured conversation or similar)

In general, the long-term recording can be segmented manually; however, some requirements should be set up to make the selected data as regular as possible, especially the length of each speech utterance:

- File length 6 to 12s per sample (segment), as common practice for P.800 listening experiments

3.1.4 Reference and degraded speech material

A minimum set of 50 samples from at least eight different talkers are required in each database where no repetition of text is allowed in this set of 50 samples. The reference files for simulated and electrical insertion conditions are excerpts that are extracted from clean spontaneous conversations, recorded at 48 kHz sample rate. For live talking conditions a human talker is speaking directly into a terminal's microphone (acoustical insertion). Potential differences in the presentation level are allowed and are part of the test conditions.

There should be a minimum of 1s of non-active speech in the sample, where non-active parts may include breathing or similar sounds. There is no need of muted segments at the start or end of the file; however, the start or end must not be located within an utterance. The speech activity according to ITU-T P.56 calculated over the entire file has to be at least 50%.

Electrical and acoustical insertions may be mixed within one database. However, there must not be a mix of common P.800 sentence pairs and excerpts of spontaneous conversations in one subjective experiment.

Level variations are included as test cases. Those level differences will be restricted to a range of +6dB to –20dB relative to a nominal level of the test application (in case of electrical recording -26dBov or equivalent). All recordings have to include a digital level as well as the sound level used in the auditory test for each database.

For all test signals delivered for fullband, a digital level of -26dBov (obtained with P.56) corresponds to the nominal presentation level (e.g. 73dB in case of diotic presentation). The actual presentation level can be directly derived from the P.56 level of the degraded signal. (e.g. a level of -34dBov corresponds to a presentation level 8dB below the nominal level).

The adjustment of the listening devices has to be done accordingly. Here the adjustment can be derived by playing out a calibration signal at -26dBov (recommended: speech like babble noise, spectrally shaped according P.50, will be made available to all proponents) over the listening device and adjusting the sound level at (A-Weighted) ERP by means of an artificial ear to the nominal level.

Note: The existing narrow-band databases were presented at approximately 79dB SPL. They can have small level variations. Usually, the digital level of -26dBov corresponds here to 79dB SPL at the ERP with monotic presentation. However, some existing data bases (SG12 Suppl. 23) are scaled such that -30dBov corresponds to 79dB SPL.

3.2 Listeners group

Normal hearing in the audio bandwidth up to 8kHz is required by the test subjects (max. hearing loss of 20dB). The subjects are split into three groups according to their age. These groups consist of subjects who are between 15 and 30 years, between 30 and 50 and above 50 years old. At least 20% of all the subjects in the experiment must fall into each group. Within each group at least 40% of the subjects must be female and at least 40% must be male.

This is a requirement for the mandatory experiments newly conducted by the proponents only. Other experiments must at least follow the requirements specified in Section 2.4.

3.3 Presentation

The digital level (ASL according to ITU-T P.56) of the signals has to be -26dBov for presentation at the nominal level. The corresponding nominal levels in the acoustical domain in the auditory tests are:

- 73 dB SPL at ERP for diotic headset tests at both ears (common case for all new experiments as described in sections 3.4 and 3.5 with headsets / diotic presentation).

For all tests in super-wideband mode the variation of the presentation level is subject to test. The allowed range is +6 to -20dB relative to the nominal level. The level will be measured in the electrical domain for simplification.

In the case of acoustical recordings, it is recommended that the same sound level is used as that produced by the terminal during acoustical capture. The desired nominal level has to be represented by a corresponding level of -26dBov as well, and level deviations are calculated related to this value. Acoustical recordings cover captures using headsets, handsets and loudspeaker phones.

The presentation of the captured files has to use diffuse-field equalized open headphones.

4 Main terms for statistical evaluation

4.1 Result accuracy and calculation of performance indicators

Each result calculated by the P.SAMD candidate is limited to four digits after the decimal point. In the case that values are averaged to gain so-called per-condition values, there is no limitation in the number of digits after the decimal point.

The precision of any calculated statistical figure is also restricted to four digits after the decimal point.

4.2 Main objectives, primary and secondary analysis

The performance of the candidate models is evaluated by using the results of subjective tests. The primary analysis is based on per-condition scores; the secondary analysis applies to 'per-sample' scores. The metrics are described in P.1401.

4.3 Analysis and performance criteria

It is agreed that the P.SAMD model will be evaluated on samples with only one distortion type as well as with multiple distortion types. Data containing multiple distortions have to be scored separately for the individual distortion types, following the procedure described in Section 2.4.

The entire evaluation procedure will be applied separately for the overall quality and for each perceptual dimension.

Statistics and objectives for P.SAMD

The evaluation is made for each experiment individually and each of the dimensions are evaluated separately by using identical statistical means.

The following procedures will be applied to each experiment and each dimension

a) Aggregation of MOS scores and predictions per condition (usually four talkers)

b) Application of a third order monotonous polynomial fitting function

c) Calculation of a per-condition rmse* as a main and single performance criterion

Objectives for training databases (min. five per set):

- Overall performance: The averaged rmse* across all experiments and dimensions/overall quality must be smaller than 0.25
- Worst case performance: No single rmse* per experiment and dimension/overall quality shall be above 0.4

Objectives for unknown, new databases (min. five per set, min. two languages per set) for P.SAMD

- Overall performance: The averaged rmse* across all experiments and dimensions/overall quality shall be smaller than 0.35
- Worst case performance: No single rmse* per experiment and dimension/overall quality shall be above 0.5

There is no experience neither in multi-dimensional testing nor in multi-dimensional prediction. In the event single performance values are not meeting the requirements, Q9/12 has to discuss and to decide how to proceed.

References

[1] ITU-T TD 12Rev1 (WP2/12) Statistical Evaluation Procedure for P.OLQA v.1.0. Rapporteur Q9/12, ITU-T SG12 Meeting, Geneva, 10-19 March 2009.

[2] ITU-T TD 90Rev4 (WP2/12) Requirement specification for P.Objective Listening Quality Assessment (P.OLQA). Rapporteur for Question 9/12, ITU-T SG12 Meeting, 22-30 May 2008.

[3] ITU-T TD 287Rev4 (GEN/12) Progress Report for Question 9/12. Rapporteur Q9/12, ITU-T SG12 Meeting, 18-27 May 2010.

[4] ITU-T Contrib. COM12-C82. Assessment of speech quality dimensions: Methodology, Experiments, Analysis. Deutsche Telekom AG, ITU-T SG12 Meeting 3-11 November 2009.

[5] ITU-T P.806 A subjective quality test methodology using multiple rating scales, Geneva, Feb 2014 .

[6] ITU-T P.808 Subjective evaluation of speech quality with a crowdsourcing approach, Geneva, June 2018.

_____