RELATIONSHIP BETWEEN ROOT CAUSES OF IMPAIRMENTS AND PERCEPTUAL QUALITY DIMENSIONS OF SUPER-WIDEBAND TRANSMITTED SPEECH

Sebastian Möller[1,2], Tobias Hübschen[3], Gabriel Mittag[1], Gerhard Schmidt[3]

[1]*Quality and Usability Lab, Technische Universität Berlin, Germany*
[2]*Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany*
[3]*DSS, Christian-Albrechts-Universität zu Kiel, Germany*
sebastian.moeller|gabriel.mittag@tu-berlin.de, gus|thu@tf.tu-kiel.de

Abstract: For diagnosing quality degradations in speech transmission and communication services, auditory degradation decompositions by quantifying perceptual quality dimensions have been established. However, it is still unclear which technical characteristics of the underlying systems are the root causes for each perceptual degradation. In order to investigate the relationship between perceptual quality dimensions and technical root causes, 8 databases of super-wideband transmitted speech samples judged regarding 3 or 4 perceptual quality dimensions have been analyzed. The results show that each perceptual dimension is triggered by a variety of (plausible) root causes. In turn, several root causes trigger more than one perceptual quality dimension. Implications of these findings for instrumentally estimating perceptual quality dimensions are discussed.

1 Introduction

Speech communication services have undergone a fundamental change within the last decade. With the introduction of IP-based transmission in the backbone network and the development of new codecs which are able to transmit speech beyond the standard 300-3400 Hz (narrowband, NB) frequency band, high-quality communication services in wideband (50-7000 Hz, WB), super-wideband (20-14000 Hz, SWB), or fullband (0-20000 Hz, FB) come to the fore. They enable a higher overall quality compared to their NB counterpart: This quality advantage has been quantified to be 29% for WB [1], and 48% for SWB [2] compared to NB, when expressed on a transmission rating scale in the range [0;100] of a popular rating model; the difference between SWB and FB turned out to be not significant in experimental studies [3].

Importantly, not only the maximum achievable quality changes, but also the character of the transmitted sound: it sounds less distant and colored compared to a direct airpath. In turn, degradations introduced by the communication system (sending terminal, circuit noise, coder, wireline or wireless transmission, decoder, receiving terminal) and the communication environment (ambient noise) might become more audible, and they might change their perceptual character. In order to ensure an optimum quality for the users, it is important for communication service providers to not only know the maximum achievable overall quality, but also to gain insights into the perceptual character of the user's experience. It is commonly expected that a perceptual analysis supports the diagnosis of sub-optimum quality, thus that perceptual characteristics pinpoint at underlying technical root causes of a degradation.

For a listening-only situation (to which this paper will be limited), Wältermann et al. [4] derived three perceptual dimensions from multidimensional analyses, which were designed to be orthogonal and valid for both NB and SWB speech: *Coloration*, *noisiness* and *discontinuity*. As the active speech level was not varied in the underlying experiments leading to these three dimensions, additional experiments have been carried out which led to a fourth dimension *sub-optimum loudness*, which is still valid from NB to SWB speech, but not necessarily

orthogonal to the other three dimensions [5]. The four perceptual dimensions were considered relevant enough to start two work items in Study Group 12 of the International Telecommunication Union (ITU-T). The proposed work items aim at predicting these four dimensions from either the input and output signal of the communication channel (reference-based model to be developed as work item P.AMD), or from the output signal alone (reference-free model to be developed under work item P.SAMD). Several model candidates have been proposed for both P.AMD and P.SAMD [5][6][7][8][9], but they have not yet been standardized, as their performance on independent test databases was not yet proven.

The mentioned perceptual quality dimensions might be triggered by different technical root causes: As an example, the transmitted audio bandwidth will have a major impact on perceived *coloration*; packet or frame loss can be expected to cause interruptions and thus *discontinuity*; circuit and ambient noise should cause perceived *noisiness*; and level alterations might provoke *sub-optimum loudness*. Whereas some of these cause-effect relationships have been identified in [5], a thorough analysis of the relationship between perceptual dimensions and technical root causes for super-wideband speech transmission is still missing, and will be the main topic of this paper. The results of such an analysis will be very helpful for developing prediction models for the P.AMD and P.SAMD work items, as the relevant root causes will be reflected in the signal(s) from which the respective perceptual dimensions shall be predicted; thus, the identification is necessary to make the respective models robust and generalizable to technically different speech communication services.

The present paper performs a thorough analysis of this relationship by means of 8 databases obtained from service providers, measurement equipment manufacturers, and research institutions. Each database contains speech samples degraded in a controlled (simulated) way, or recorded in life networks (with only partial control over some of the technical characteristics); each speech sample has been judged on 3 or 4 of the mentioned perceptual quality dimensions, according to a standardized method described in the Requirement Specification of P.AMD [10]. The databases are described in Section 2, and the relationships between technical root causes and each perceptual dimension are analyzed for each database in Section 3. Section 4 summarizes the results and interprets them regarding the development of diagnostic tools, and the prediction models for P.AMD and P.SAMD. Finally, Section 5 draws some conclusions and outlines future work.

2 Databases

The 8 databases have been built at different points in time, and partially for different purposes, resulting in slightly different characteristics. Two of them (DTAG 1 and DTAG 2) stem from the identification of the perceptual dimensions, and were collected when the dimension *sub-optimum loudness* was not yet identified; thus, these databases contain only 3 perceptual dimension judgments, and they only address NB and WB speech samples. Five others (DTAG 3, Orange, and the 3 SwissQual databases) were part of the development process for a prediction model for overall quality in ITU-T SG12 (P.OLQA competition), and were later annotated regarding their perceptual dimensions. The final database (TUBDIS) was used for the development of P.AMD-type models. The databases are available to interested parties under a particular license; details can be obtained from the Rapporteur of Q.9/12 of ITU-T Study Group 12.

Each database consists of the degraded speech files with a brief description of the processing conditions (either simulated degradations, or life speech calls recorded in a network), and with 3 or 4 averaged ratings for the perceptual dimensions on a scale in the range [1;5]. While acknowledging the problems incurring with averaged ratings due to the non-Gaussian distribution of the individual ratings, only the averages were available to us and could thus be

used in the analysis. The ratings were obtained from native listeners of the respective language, who listened to each file in a sound-insulated environment over headphones, and judged the perceptual dimensions according to the procedure developed by Wältermann [11] and described in [10]. Other relevant characteristics of each database are summarized hereafter.

DTAG 1-3: All DTAG databases were collected in German language. DTAG 1 and 2 contain source material from 2 talkers (1m, 1f), using 12 files per condition. Each file was rated by 3-4 (DTAG 1) or 4-5 (DTAG 2) listeners, resulting in 40/48 votes per condition. DTAG 1 contains 66 test conditions, DTAG 2 both NB and WB speech. DTAG 1 was mostly degraded by codecs, incl. codec tandems, whereas DTAG 2 showed a larger variety of degradations, e.g. ambient noise of different level and character, mostly in conjunction with different coding algorithms. In addition, DTAG 2 contains packet loss, different bandpass filters, as well as signal-correlated noise generated by a Modulated Noise Reference Unit, MNRU [12]. DTAG 3 has more variety regarding talkers (2m, 2f), and was built with 4 files per condition, each file rated by 20 listeners. The 54 test conditions range from NB to SWB, and contain realistic (electrical or acoustic) recordings from life networks, some involving background noise, as well as some artificially-degraded conditions with additive or MNRU-generated noise, bandpass filters, level attenuations, as well as time clipping.

Orange 1: The Orange 1 database was collected by Orange Labs, Lannion, in the French language. It contains material from 4 talkers (2m, 2f) which was degraded by 56 circuit conditions, each file rated by 18 listeners with respect to all four perceptual dimensions. Circuit conditions were WB and SWB, and include background and circuit noise of different types and level, signal-correlated noise generated by an MNRU, recordings through terminal equipment, attenuations, temporal clipping, packet loss, and time warping.

Swissqual 1, 501, 502: All SwissQual databases use recordings from 4 Swiss-German speakers (2m, 2f) as source material, and have ratings along all four perceptual dimensions. SwissQual 1 contains 4 source files per circuit condition, with 24 ratings per file. The 30 circuit conditions range from NB to SWB, and include anchor conditions which are expected to trigger one perceptual dimension each, codecs, ambient noise, time clipping, loudness loss, as well as many life recordings. SwissQual 501 uses only 2 source files per circuit conditions, with again 24 ratings per file. The 50 circuit conditions range from NB to SWB and include anchor conditions, codecs, packet loss, background noise, as well as acoustic recordings and life calls. SwissQual 502 is very similar to SwissQual 501, with slightly different but comparable circuit conditions.

TUBDIS: The TUBDIS database was recorded at TU Berlin, Germany, and uses Swiss-German source files from the SwissQual databases (2 talkers, 1m and 1f). Each of the 20 conditions was rated by 35-41 listeners; circuit conditions are all SWB and include anchor conditions, codecs, and packet loss. All 4 perceptual dimensions were rated by the listeners.

3 Analysis

The average per-condition judgments of all perceptual dimensions of all databases will now be analyzed with respect to the technical root causes, which are the basis of degradations for any of the perceptual dimensions. Exemplary results for the DTAG 1 database, indicating the average per-condition judgments for each available perceptual dimension (in this case noisiness, discontinuity and coloration), are shown in Fig. 1. In order to use a common criterium for what can be considered as a “degradation”, and in order to cope with the multitude of databases and conditions, we extracted circuit conditions for which the average perceptual dimension rating is lower than 3.0 on the 5-point judgment scale. For some circuit conditions, only imprecise descriptions are available, namely when source speech files were degraded by

life calls and/or acoustic recordings in real-life environments. As a consequence, technical root causes cannot always be extracted. For all other circuit conditions, the known root causes will be listed.

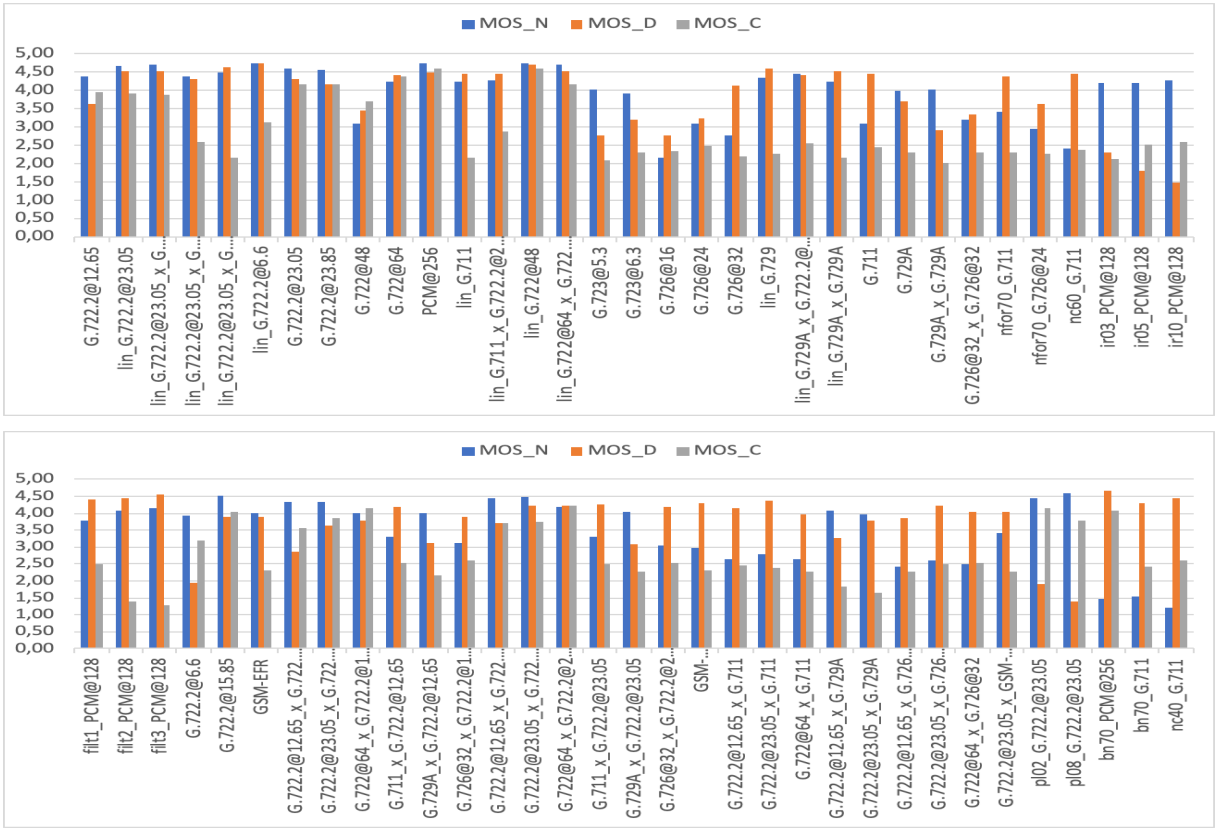

Figure 1 – Average per-condition ratings for noisiness (MOS_N), discontinuity (MOS_D) and coloration (MOS_C) for the DTAG 1 database. Condition identifiers indicate codec@bitrate (in kbit/s), codec tandems (x), circuit noise on the channel (ncl) or at the receiving side (nforl) of level l dBm0p, background noise (bnl) at the sending side of level l dB(A), random packet loss (pln) or interruptions (irn) of n %, or different bandpass filters (filtn).

For **DTAG 1**, degradations of perceptual dimensions occurred in the following cases:

- *Coloration*: All cases of NB coding, namely G.711, G.729A, GSM-EFR, G.723 at 5.5 or 6.3 kbit/s (G.723@5.5 and G.723@6.3), G.726@24 and G.726@32, as well as tandems of these with other codecs. In addition, several bandpass filter conditions provoked coloration ratings lower than 3.0, as well as combinations of NB coding and background noise or circuit noise.
- *Noisiness*: Most background and circuit noise conditions, as well as G.726@16, G.726@24 (partially) and G.726@32. Noisiness was also provoked by the codec tandems G.722.2*G.711 and G.722*G.711.
- *Discontinuity*: Is caused by codecs G.723@5.3, G.726@16, and G.722.2@6.6, as well as the codec tandems G.729A*G.729A and G.722.2@12.65* G.722.2@12.65. As expected, also 3, 5 and 10% of interruptions caused discontinuity.

DATG2: Perceptual dimension degradations were observed for the following root causes:

- *Coloration*: NB coding by G.711 or G.729A, as well as the combination of a clean NB PCM channel with signal-correlated MNRU noise.
- *Noisiness*: Is caused by background noise recorded from a car (≥55 dB(A)) or in a pub (≥70dB(A)). In case of NB transmission (using the G.711 codec), noisiness ratings be-

low 3.0 are already observed from lower background noise levels (≥45 dB(A)) onwards. In addition, also an MNRU with SNR=10 dB, in some cases also with SNR=20dB, causes noisiness.

- *Discontinuity*: Is caused by packet loss rates as low as 2%, if the lost packets are not masked by noise. In the presence of noise, packet loss rates of 4 or 8% caused discontinuity. In case of G.722.2 coding, or G.722 with lower bitrates, packet loss rates of 1% were sufficient to cause discontinuity. Interestingly, also MNRU with SNR=10 dB and partially also SNR=10 dB caused discontinuity.

DATG3: For this database, also mixed life recordings provoke dimension ratings below 3.0. However, as the exact technical settings of these conditions are unknown, only the simulated conditions will be discussed regarding possible root causes:

- *Coloration*: Is caused by filtering with a 500-2500 Hz bandpass filter, by applying a modified IRS sending characteristic according to [13], as well as by several NB codecs, triple NB codec tandems, as well as by the AMR-WB codec operating at the low 6.6 kbit/s bitrate. In addition, MNRU with SNR=10 dB in combination with background noise caused coloration.
- *Noisiness*: As expected, this is caused by background noise, MNRU-type noise, as well as circuit noise. In addition, also AMR-NB coding caused noisiness.
- *Discontinuity*: Is only observed for 20% time clipping.
- *Loudness*: In this database, also sub-optimum loudness was rated. Values lower than 3.0 on this dimension are observed for level reductions by -20dB, for triple codec tandems, as well as for the G.729.1@14 codec.

Orange1: The following root causes can be identified:

- *Coloration*: Is caused by MNRU with SNR=10 dB, by 500-2500 Hz bandpass filtering, as well as by applying a modified IRS send side filter. In addition, acoustic recordings with a hands-free terminal and a handset cause coloration. As expected, most NB conditions are rated with coloration <3.0. Interestingly, also 2% amplitude overload as well as some conditions with >2% time clipping are perceptually degraded by coloration.
- *Noisiness*: Is caused by MNRU-type noise, background noise (car, street, babble), as well as by a circuit noise floor.
- *Discontinuity*: Is caused by time clipping ≥ 10%, by gain variation, as well as by wireless frame errors of 0.2%. In addition, the combination of 2% temporal clipping and car noise cause discontinuity.
- *Loudness*: Sub-optimum loudness is caused by a level of -20dB, as well as gain variation and by applying an active gain control (AGC) device.

SwissQual 1: All SwissQual databases contain many life recordings for which the root causes are unknown. Thus, only the conditions with known technical characteristics are summarized in the following:

- *Coloration*: Is caused by 500-2500 Hz bandpass filtering, as well as by the NS codec.
- *Noisiness*: Is caused by (Hoth-type) noise with SNR= 16…20 dB, as well as by an MNRU with SNR=10 dB.
- *Discontinuity*: Is caused by time clipping in the range 2…20%.
- *Loudness*: Sub-optimum loudness is observed for a level attenuation of-10 dB or more.

SwissQual 501: The following perceptual dimension degradations are observed:

- *Coloration*: Caused by 500-2500 Hz bandpass filtering, modified IRS send side filtering [13], AMR-NB in Mode 5 (7.95 kbit/s), as well as by an MNRU with SNR=10 dB. Interestingly, also coding with AMR-WB in Mode 0 (6.6 kbit/s), as well as EVRC-B coding, cause coloration.
- *Noisiness*: Is caused by noise with SNR<20dB, as well as by MNRU-type noise with SNR=10dB. Also many acoustic life recordings contain noisiness.
- *Discontinuity*: Is caused by 20% time clipping, as well as by packet loss. The life recording conditions marked as "bad channel" also provoke discontinuity.
- *Loudness*: Is rated as sub-optimum if the speech level is attenuated by more that -16dB.

SwissQual 502:

- *Coloration*: A variety of conditions provoke coloration, including an MNRU with low SNR, some narrow bandpass filters, as well as the codecs NS, Windows Media Audio, AMR-NB@12.2, EVRC-B, QCELP, G.711, EVRC-A, and AMR-NB. Recordings via an acoustic terminal in sending direction also contain coloration.
- *Noisiness*: Is perceived in all noise conditions (except MNRU with SNR=25dB), in many life calls, and in case of acoustic noise at the receiving end if the overall level is not attenuated (exception: video call AMR-NB with -16dB attenuation plus acoustic noise).
- *Discontinuity*: Is only caused by artificial time clipping and amplitude clipping.
- *Loudness*: Is rated as sub-optimum if the level is attenuated by -8dB in the presence of noise, or by -10 dB without noise.

TUBDIS

- *Coloration*: Is observed for bandpass filtering in the 500-2500Hz and 100-5000 Hz frequency range, for packet loss with 3% (partially) and 6%.
- *Noisiness*: Is observed for noise with a level higher than 12 dB, as well as for 6% packet loss.
- *Discontinuity*: Caused by time clipping and packet loss ≥3%.
- *Loudness*: Is rated sub-optimally for attenuations of -10dB.

4 Discussion

Summarizing the analysis results of the 8 very diverse databases, it seems that most of the root causes of impairments seem to be in line with the perceptual quality dimensions they trigger, namely:

- *Coloration*: In a WB and SWB experimental listening context, NB codecs cause coloration, especially when they come at low bitrate or in tandems; NB-filtered MNRU also seems to cause coloration; lowpass and tight (e.g. 500-2500 Hz) bandpass filtering, as well as modified IRSsend characteristics and acoustic terminals degrade this perceptual dimension; in contrast to our expectation, in some cases also packet loss, clipping and overload can cause coloration.
- *Noisiness*: Is caused by MNRU-type noise, circuit noise, and background noise; also the G.726 codec (at low bitrates) and in some cases AMR-NB cause noisiness; interestingly, sometimes also high packet loss is perceived as noisy.

- *Discontinuity*: Is caused by packet loss of moderate and high rate, if not masked by noise; in addition, interruptions, time clipping, amplitude clipping gain variations, and wireless errors cause discontinuity; unexpectedly, NB codecs at low rate or in tandem (e.g. G.726@16, G.723@5.3, G.729A*G.729A), and even MNRU cause discontinuity.
- *Loudness*: Is caused by moderate to strong attenuation, by gain variation, and by AGC.

An important finding is that there are several root causes which cause degradations on *more than one* perceptual dimension. Examples of these causes are low-bitrate codecs which cause coloration and discontinuity, the G.726 codec which causes coloration and noisiness, packet loss which causes discontinuity and coloration, as well as MNRU-type noise which causes noisiness, discontinuity and coloration. In turn, there is commonly a wide range of root causes behind each perceptual dimension. This proves that a 1-to-1 relationship between technical root causes and perceptual dimensions is not possible, as both directions of the relationship are ambiguous.

A second important finding are masking effects, in the sense that discontinuity (e.g. caused by moderate rates of packet loss) is sometimes masked by noise, or that noisiness may be masked by loudness loss. This illustrates that technical root causes – when occurring in combinations – have to be considered together for defining the respective perceptual dimension degradations.

Despite the ambiguity in the relation between technical causes and perceptual dimensions, it seems that some knowledge on the technical set-up of the system will be helpful to identify root causes from the perceptual dimensions. As an example, if the codec of a speech communication channel is known, then it can be anticipated which perceptual dimensions it will degrade. Any other remaining perceptual degradations which are either (subjectively) measured or (instrumentally) predicted for such a channel can then be associated with other root causes. In case of life channel recordings, some dominant effects can be identified with the help of the perceptual analysis; however, this could not be fully proven, as the technical settings of the life channels were unknown for our databases.

A shortcoming of the presented analysis is that one fixed threshold (3.0) was used to define a "degradation" along all perceptual dimensions. This way, information which is contained in the continuous averaged judgments of the test participants gets lost, and all dimensions are considered to be equally important contributors to overall quality. We used this fixed threshold to keep the analysis method constant throughout the large number of conditions and databases, but acknowledge that the continuous ratings may contain additional diagnostic information which can be of relevance for network operators. In practice, different thresholds may be defined for each perceptual dimension, reflecting a weighting of each degradation towards overall quality, see e.g. [14].

The listing of root causes for each of the perceptual dimension given above will help to set up estimators for individual perceptual dimensions. For example, an estimator for coloration would take profit if it does not only work on the spectral (bandwidth) characteristics of the signal, but is also able to identify codecs at low bitrate [15]. Or, a noisiness estimator might also need to identify the G.726 codec, especially if working at lower bitrates. Or, a discontinuity estimator does not only need to find gaps (packet loss, time clipping), but also amplitude clipping and level overload.

5 Conclusions and future work

We presented a qualitative analysis of 8 databases containing NB to SWB transmission conditions degraded by a variety of technical toot causes. While acknowledging that a quantitative analysis would have been desirable, this was prohibited by the databases: Only mean ratings

were available to the authors, and the conditions differed in many respects, making a quantitative cross-database comparison impossible. The results of our analysis show that there is no 1-to-1 relationship between technical causes and perceptual dimensions, but that ambiguity exists in both directions. Nevertheless, several root causes which regularly coincide with perceptual dimension degradations could be identified. These root causes should be identified and quantified from the signal characteristics in order to provide valid perceptual dimension predictions in the P.AMD and P.SAMD frameworks. In addition to the development of such predictors, we consider an analysis of life recorded data as a worthwhile target for future work, in order to check the applicability of the diagnostic framework in practical monitoring scenarios.

This work has been funded by the Deutsche Forschungsgemeinschaft, DFG, grants MO 1038/22-1 and SCHM 1507/10-1. We thank all parties who have contributed databases.

Literatur

[1] ITU-T REC. G.107.1: *Wideband E-model*. Int. Telecomm. Union, CH-Geneva, 2015.

[2] ITU-T CONTR. SG12-C.297: *Draft Recommendation G.107.2 'Fullband E-model'*. Source: Deutsche Telekom AG, NTT, Japan, Orange, France, Int. Telecomm. Union, CH-Geneva, Nov. 2018.

[3] ITU-T CONTR. SG12-C.260: *Fullband Extension of R-value*. Source: NTT, Japan, Int. Telecomm. Union, CH-Geneva, Nov. 2018.

[4] WÄLTERMANN, M., A. RAAKE, S. MÖLLER: *Quality Dimensions of Narrowband and Wideband Speech Transmission*, *Acta Acustica united with Acustica* 96(6), pp. 1090-1103, 2010.

[5] CÔTÉ, N.: *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Springer, 2011.

[6] ITU-T CONTR. SG12-C.303: *P.AMD Set A Updated Performance Results of the Noisiness Dimension*. Source: Deutsche Telekom AG, Opticom GmbH, Int. Telecomm. Union, CH-Geneva, Nov. 2018.

[7] ITU-T CONTR. SG12-C.42: *First Possible P.SAMD Indicators for the Estimation of Coloration*. Source: Deutsche Telekom AG, Int. Telecomm. Union, CH-Geneva, 2017.

[8] ITU-T CONTR. SG12-C.43: *First Possible P.SAMD Indicators for the Estimation of Noisiness*. Source: Deutsche Telekom AG, Int. Telecomm. Union, CH-Geneva, 2017.

[9] ITU-T CONTR. SG12-C.300: *P.SAMD Update of Ongoing Work*. Source: Deutsche Telekom AG, Int. Telecomm. Union, CH-Geneva, Nov. 2018.

[10] ITU-T CONTR. COM 12-C195: *Draft Requirement Specification for P.AMD (Perceptual Approaches for Multidimensional Analysis*. Source: Deutsche Telekom AG, Int. Telecomm. Union, CH-Geneva, Jan. 2011.

[11] WÄLTERMANN, M.: *Dimension-based Quality Modeling of Transmitted Speech*. Berlin, Heidelberg: Springer, 2012.

[12] ITU-T REC. P.810: *Modulated Noise Reference Unit (MNRU)*. Int. Telecomm. Union, CH-Geneva, 1996.

[13] ITU-T REC. P.830: *Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs*. Int. Telecomm. Union, CH-Geneva, 1996.

[14] WÄLTERMANN, M., A. RAAKE, S. MÖLLER: *Analytical Assessment and Distance Modeling of Speech Transmission Quality*. In *Proc. 11th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech 2010)*, 26-30 Sept., JP-Makuhari, 2010.

[15] HÜBSCHEN, T., G. SCHMIDT: *Bitrate and Tandem Detection for the AMR-WB Codec with Application to Network Testing*. In *Proc. EUSIPCO*, IT-Rome, 2018.