Comparison of Transmission Quality Dimensions of Narrowband, Wideband, and Super-Wideband Speech Channels

Sebastian Möller[1], Friedemann Köster[1]
[1]Quality and Usability Lab, Telekom Innovation Labs
Technische Universität Berlin, Germany
sebastian.moeller@telekom.de,
friedemann.koester@telekom.de

Laura Fernández Gallardo[2,1], Michael Wagner[2,3,4]
[2]University of Canberra, [3]Australian National University,
[4]National Centre for Biometric Studies, Australia
laura.fernandezgallardo@canberra.edu.au,
michael.wagner@canberra.edu.au

***Abstract*— It is commonly acknowledged that the introduction of wideband and super-wideband speech transmission in Voice-over-IP leads to an improved overall quality compared to traditional narrowband telephony. However, beyond overall quality, dimensions such as coloration, continuity, noisiness, loudness, human speaker identification ability, as well as automatic speech and speaker identification performance may benefit from the augmented speech transmission bandwidth. In this paper, we review data on all of these quality and performance dimensions and classify them along a recommended "transmission rating scale" into eight service quality classes which are useful for transmission network planning. Applications of the results are discussed for different human-to-human and human-machine application scenarios.**

Keywords—speech quality, quality dimensions, speaker identification, Voice-over-IP, wideband

I. INTRODUCTION

With the introduction of Voice-over-IP (VoIP), it is easily possible to transmit speech beyond the classical 300-3400 Hz transmission bandwidth. Whereas in the Integrated Services Digital Network (ISDN) this was only possible by connecting multiple circuit-switched channels, it now suffices to agree on a joint codec in the sending and the receiving terminals to enable a wider speech bandwidth. Depending on the speech codec used, the actual transmission bandwidth may actually be lower than in a narrowband transmission scenario. For example, the AMR-WB codec, according to ITU-T Rec. G.722.2, allows transmitting with bandwidths between 6.6 and 23.85 kbit/s, and the subband ADPCM, according to ITU-T Rec. G.722, at bitrates between 48 and 64 kbit/s, compared to a logarithmic PCM with 64 kbit/s which is the standard for ISDN telephony.

Despite the lower bitrate, transmission quality is commonly assumed to increase with increasing speech bandwidth. For example, the overall quality of wideband (WB, 50-7000 Hz) transmission was found on an average to be 29% higher than that of a narrowband (NB, 300-3400 Hz) transmission channel [1][2]. For super-wideband (SWB, 50-14000 Hz), Wältermann et al. [3] found an increase of the overall quality of up to 79%, when expressed on a common quality scale. These values need to be slightly reduced in case that transmission degradations apart from the channel bandwidth limitations are introduced, e.g. by speech codecs, packet loss, background noise, signal processing, or alike. Still, it is to be expected that the overall transmission quality increases on average.

Quality has been described as a multidimensional "event", which consists of several perceptual dimensions (so-called quality features [4]). These features can be extracted by performing multidimensional analyses on transmitted speech files, e.g. by judging their perceptual similarity and performing a subsequent multidimensional scaling (MDS), or by judging each file on a semantic differential (SD) scale and performing a subsequent principal component analysis, see e.g. [5] for a description of such techniques. A number of such features are collected in the Qualinet White Paper on Quality of Experience [4]. Analyzing narrowband and wideband channels with respect to the listening situation (to which this paper will be limited), Wältermann et al. identified three dimensions, which were termed discontinuity, noisiness, and coloration [6]. A fourth dimension was added later reflecting the non-optimum loudness [5]. Other authors [7] extracted more (up to seven) such dimensions, separating discontinuity into slowly-varying and fast-varying fluctuations, and coloration into low-frequencies absent and high-frequencies absent types of coloration.

In practical application scenarios, however, perceptual dimensions as extracted by MDS or SD are not always easily interpretable. Instead, telecommunication engineers and decision-takers need to have figures of merit of practical relevance, such as intelligibility scores, speaker identification scores, or alike. In the best case, such scores are extracted with the help of subjective experiments with human test participants, carried out under controlled laboratory conditions. To avoid the effort required for carrying out such experiments, figures are sometimes also estimated with the help of instrumental (or so-called "objective") prediction models. When carefully chosen, models may adequately estimate scores for a specific transmission scenario and target score. To date, only models for predicting overall quality, perceptual quality dimensions, and (partially) intelligibility are known for telephony applications.

978-1-4799-5255-7/14/$31.00 ©2014 IEEE

Figures of merit are also relevant for machine recognition applications, i.e. when the transmitted speech is input to an automatic system such as automatic speech recognition, automatic speaker recognition, automatic emotion recognition, or alike. Such systems are spreading in call centers to automate the call flow and adequately address customer wishes. In that case, figures of merit are mostly specified in terms of recognition accuracies or error rates.

For transmission service planners, it is difficult to estimate what improvement an increased speech transmission bandwidth could bring for a specific application. Planners prefer to have a unified figure of merit, associated with a specified service class, which may justify decision-taking with respect to the speech transmission bandwidth. In the past, transmission planners relied on figures of merit on a so-called "transmission rating scale" which has been defined by the E-model, a computational model for transmission planning, in ITU-T Rec. G.107 [8]. This scale ranges from 0 (worst imaginable quality) to 100 (optimum quality of a narrowband connection), and requirements for service quality classes on the basis of the transmission rating are given in ITU-T Rec G.109 [9]. The scale has recently been extended towards wideband channels, providing a maximum transmission rating of 129 and leaving the narrowband range of the scale untouched [10]. For SWB channels values above 130 have been suggested but have not yet been confirmed.

In this paper, we will provide an attempt to quantify the advantage of wideband speech transmission for different quality dimensions and applications, on a common scale, and define service quality classes on that basis. Our approach is based partially on newly-processed and partly on pre-published data, which we will classify on the basis of the figures of merit for different criteria and express in terms of ranges on the transmission rating scale. This way, we hope to keep the simplicity of the transmission rating approach, while providing more in-depth information for some typical transmission scenarios.

The transmission scenarios and the quality criteria which are relevant for different application scenarios are briefly outlined in Section 2. Section 3 analyzes data on each of the criteria which can be used for the approach. Section 4 describes the classification results. Section 5 discusses practical use cases for the approach and concludes with an outlook on future work.

II. Transmission Scenarios and Criteria

In modern speech transmission scenarios, a number of degradations can be expected to be relevant for performance and quality of different services, apart from the speech bandwidth to-be-transmitted (see Fig. 1). These include:

- Coding distortions
- Packet loss or discard, and effects of imperfect packet loss concealment
- Degradations due to the sending and receiving terminal, including the electro-acoustic characteristics of the transducers, and effects of imperfect signal-processing equipment (voice activity detection, noise reduction, echo cancellation, etc.) integrated in the terminal
- Attenuation
- Background and circuit noises (remaining after imperfect noise reduction)
- Overall delay
- Talker and listener echoes

Amongst these, attenuation seems to be of minor importance for modern transmission scenarios. Noises can have different effects on human-perceived quality, depending on how they are evaluated: Whereas the overall quality may be significantly improved by noise reduction, speech intelligibility might be considerably degraded in turn, as an overly effective noise reduction algorithm might remove relevant spectral energy contributions from the speech signal. This is reflected by the three-fold methodology which is currently recommended by the International Telecommunication Union (ITU-T) for estimating speech quality in the presence of noise; see ITU-T Rec. P.835 [12]. In addition, different types of noises have different impacts on speech quality [13]. This renders the comparison of noise-originated degradations with other quality degradations quite difficult.

The effects of delay and echo can only be evaluated in a conversational situation, and heavily depend on the conversation scenario being considered (see e.g. the discussions in [8]). Thus, in our comparison we will concentrate on the first three types of degradations, as these can be considered for different quality aspects and service usage scenarios alike.

The most typical usage scenario of a telephone channel is obviously a human-to-human telephone conversation. This scenario can be evaluated in terms of the overall quality, as well as with respect to the perceptual quality dimensions mentioned above. In addition to quality, speech intelligibility plays a major role, and this can be expected to depend heavily on the transmitted speech bandwidth. In classical application scenarios such as for room-acoustic transmission and public address systems, the Speech Transmission Index (STI) is a widely-used and easily applicable method to estimate intelligibility on a syllable, word, or sentence level [14][15]. However, it is not recommended for measuring intelligibility for telephone (and especially VoIP) channels with low-bitrate codecs and packet-loss degradations. Interestingly, an intensive

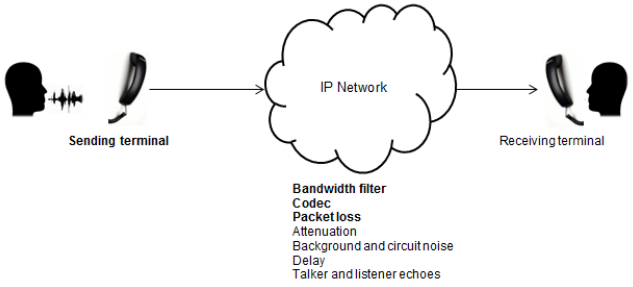

Fig. 1. Transmission scenario and principal impairments

TABLE 1: QUALITY AND PERFORMANCE METRICS FOR DIFFERENT CRITERIA, USING DIFFERENT INSTRUMENTAL MODELS OR IDENTIFICATION/RECOGNITION TASKS. FIRST BLOCK: NB SCENARIOS; SECOND BLOCK: WB SCENARIOS; THIRD BLOCK: SWB SCENARIOS.

Criterium				Speech Quality									Speaker identification				Autom. Recog.	
Task or instrumental model				POLQA	WB-PESQ	DIAL					WB E-model*		Human speaker identification (%)				Speaker	Speech
Sending Device	Band-width	Codec (bandwidth)	Packet loss (%)	POLQA MOS	WB-PESQ MOS	DIAL MOS-C	DIAL MOS-D	DIAL MOS-N	DIAL MOS-L	DIAL MOS	R	MOS	Word	Sent.	Parag.	Sent. Start	EER (%)	WA (%)
No	NB	No	0	3.42	3.93	3.00	4.31	4.34	4.02	3.34	94.2	3.74	-	-	-	-	3.41	-
No	NB	G711 (64)	0	3.59	3.27	3.07	4.30	4.23	4.43	3.24	93.2	3.70	56.65	89.18	93.75	-	4.29	92.1
No	NB	GSMEFR (12.2)	0	3.31	2.91	2.96	4.16	4.29	4.40	3.12	88.2	3.52	54.25	89.66	94.47	-	5.54	92.5
No	NB	AMRNB (4.75)	0	2.58	2.17	2.75	3.32	4.28	4.38	2.84	-	-	47.76	84.37	93.03	-	8.21	89.9
Handset	NB	G711 (64)	0	3.44	-	2.99	4.10	4.12	4.49	2.89	93.2	3.70	-	-	-	67.81	-	-
Handset	NB	G711 (64)	5	2.93	-	2.87	2.19	4.00	4.49	2.36	42.1	1.72	-	-	-	58.75	-	-
Handset	NB	G711 (64)	10	2.32	-	2.72	1.11	3.88	4.49	1.99	26.9	1.28	-	-	-	60.63	-	-
Handset	NB	G711 (64)	15	1.98	-	2.76	1.11	3.80	4.50	1.97	19.4	1.12	-	-	-	56.25	-	-
Conf. phone	NB	G711 (64)	0	2.19	-	2.71	1.44	3.73	4.04	1.98	-	-	-	-	-	60.31	-	-
Headset	NB	G711 (64)	0	3.17	-	2.97	4.28	4.05	3.83	2.87	-	-	-	-	-	66.88	-	-
Mobile	NB	AMRNB (12.2)	0	2.95	-	2.88	2.91	4.06	4.44	2.57	-	-	-	-	-	63.13	-	-
No	WB	No	0	4.13	3.18	4.08	4.31	4.27	4.13	3.72	128.8	4.50	-	-	-	-	1.46	98.2
No	WB	G722 (64)	0	4.15	3.93	4.01	4.31	4.24	4.44	3.64	115.8	4.33	66.75	93.99	95.67	-	1.80	98.2
No	WB	AMRWB (23.05)	0	3.98	3.78	3.91	4.31	4.21	4.43	3.78	127.8	4.49	67.31	94.95	96.39	-	2.86	98.1
No	WB	AMRWB (12.65)	0	3.66	3.41	3.86	4.31	4.22	4.43	3.84	115.8	4.33	-	-	-	-	2.52	-
Handset	WB	G722 (64)	0	4.00	-	3.58	4.29	3.95	4.17	3.77	115.8	4.33	-	-	-	75.00	-	-
Handset	WB	G722 (64)	5	2.48	-	3.26	1.83	3.73	4.20	2.28	75.3	3.01	-	-	-	73.75	-	-
Handset	WB	G722 (64)	10	2.13	-	3.05	1.19	3.62	4.14	2.01	61.5	2.46	-	-	-	72.50	-	-
Handset	WB	G722 (64)	15	1.85	-	3.06	1.03	3.21	4.16	1.87	54.7	2.18	-	-	-	66.88	-	-
Conf. phone	WB	G722 (64)	0	2.65	-	3.67	1.41	3.68	4.20	2.27	-	-	-	-	-	72.19	-	-
Headset	WB	G722 (64)	0	4.08	-	3.97	4.31	4.06	4.07	3.61	-	-	-	-	-	80.31	-	-
Mobile	WB	AMRWB (12.65)	0	3.45	-	3.82	2.94	3.94	4.44	2.83	-	-	-	-	-	76.88	-	-
No	SWB	No	0	4.52	-	4.07	4.31	4.26	4.17	4.11	-	-	-	-	-	-	-	-
Headset	SWB	G722.1C (32)	0	3.55	-	3.89	4.13	4.01	4.07	3.78	-	-	-	-	-	77.19	-	-
Headset	SWB	G722.1C (48)	0	3.75	-	3.93	4.21	4.07	4.08	3.79	-	-	-	-	-	77.19	-	-

*: As the WB-E-model does currently only handle handset sending devices without considering further degradations, the no device and handset conditions have been assumed to be the same for this model.

literature survey showed that few data seems to be available determining speech intelligibility for modern telephone channels; classical papers address the contributions of different frequency components to intelligibility [16][17], but they do not consider modern transmission equipment in the channel [18]. The lack of data for this criterion might also be due to missing agreement on a common and valid evaluation method for telephone-speech intelligibility. The ITU-T has recently opened a work item intending to develop such a method [19]. As long as the discussion is ongoing, we refrain from using this (otherwise very relevant) criterion in our comparison.

Another important criterion is the extraction of paralinguistic information about the conversation partner, such as his/her identity, age, gender, emotional state, personality, etc. Whereas this is a naturally-occurring phenomenon in human conversations, it is rarely being taken into account when planning transmission services. From the wealth of paralinguistic information available, we will limit our analysis to the speaker's identity as a primary criterion, leaving other information for further study.

Both linguistic and paralinguistic information is used increasingly by automatic classifiers that are part of speech-technology-based phone services, e.g. for information, reservation or transaction tasks (booking systems, online banking, etc.). As exemplary criteria, we use the ability to extract linguistic content (via automatic speech recognition) and to identify the speaker (via automatic speaker recognition).

In summary, we will use the following criteria and performance metrics for our data analysis:

- Overall quality: Mean Opinion Score (MOS) estimated either with the current standard POLQA [20] or with the long-standing previous standard WB-PESQ [21]. Whereas the first is applicable to all NB, WB and SWB channels, the latter only addresses NB and WB channels. MOS values are estimated on a joint scale in the range [1;4.5], leading to slightly lower values for NB.
- Perceptual quality dimensions: These are estimated with the help of the DIAL model [22] which has proven to pro-vide reliable estimates for the dimensions coloration, discontinuity, noisiness and loudness, as well as an additional overall quality index (DIAL-MOS), all on the same scale [1;4.5]
- Transmission planning indices: The ITU-T recommends the use of the E-model [8] in the NB case, and the WB-E-model [10] for WB channels; both models provide estimations on the transmission rating scale, which can be transformed to the MOS scale representing a WB situation using the formula given in [10].
- Human speaker identification: Here we use rates of correctly identified words, sentences or (starts of) paragraphs.

- Automatic speaker recognition: We use the Equal Error Rate (EER) as a handy criterion of speaker identification performance.
- Automatic speech recognition: The Word Accuracy (WA) is used as a performance metric.

These criteria will be analyzed for databases collected in 4 studies, described hereafter.

III. Data Analysis

In order to obtain comparable figures for the different criteria across different channels (bandwidth, codecs, sending devices, packet loss), human-labeled databases collected in two studies (Study 1 and Study 2) were available to us. They have been collected for human speaker identification experiments at TU Berlin, and the collected speech files were also used to determine indices of overall quality and of different quality dimensions, using the POLQA, WB-PESQ and DIAL models. In addition, we carried out an automatic speaker verification experiment on standard databases (Study 3), and took literature data for automatic speech identification (Study 4).

A. Databases

Study 1 contains original recordings (words, sentences and paragraphs) from 16 speakers (8m, 8f) made with a high-quality microphone in an acoustically isolated room, and distorted with the help of standardized software [23] to simulate channel filters and codecs. The distorted files were judged in a speaker-identification experiment by 26 listeners who were familiar with the speakers, see [24] for details.

Study 2 contains extracts from the clean recordings of Study 1 (beginnings of a paragraph), but played back via a head-and-torso simulator and re-recorded with different sending interfaces (handset telephone, conference phone, headset, mobile phone), again distorted with different codecs, and applying different rates of random packet loss. These files were judged by 20 listeners familiar with the speakers, see [25] for details.

For the transmission planning indices, the settings of each transmission channel were used with the NB or the WB version of the E-model, leading to a transmission rating R in the range [0;100] for NB and [0;129] for WB channels.

For the automatic speaker verification experiments in Study 3, 5 databases (TIMIT Acoustic-Phonetic Continuous Speech Corpus, Resource Management Corpus 2.0 Part 1, North American Business News Corpus 1, Wall Street Journal Continuous Speech Recognition Phases I and II) were used, extracting 20 mel-scale cepstral coefficients (MFCC) together with a log energy feature using a 25ms Hamming window with 10ms frame shift and the corresponding derivative (delta) coefficients (total of 63 coefficients). Background training was performed with an i-vector extractor trained with data of the same distortion as the test utterances, and enroll/test data consisted of the test partition of TIMIT not used for training, concatenating 5 utterances for enroll, and using cosine distance scoring. Details on this study are given in [26].

For the automatic speech recognition in Study 4, literature data from [27] was used. The authors of this study employed sentences from the TIMIT dataset transmitted through various codecs as speech material. Table 1 contains the quality and performance figures from all 4 studies, grouped according to the transmission characteristics of the respective channels.

B. Analysis for Narrowband Channels

In the NB case, optimum overall quality is reached with either no or a handset terminal, and with no or the G.711 log-PCM coding. Instrumental models like POLQA seem to be optimized towards handset listening and G.711 coding, as they indicate a higher overall quality for this condition than without coding. DIAL shows that the G.711 codec produces a little discontinuity, but only to a minor extent, leaving an overall positive picture. Quality slightly degrades on the coloration dimension when codecs such as AMR-NB and GSM-EFR are used and severely on the discontinuity dimension when packet loss is introduced. For a packet loss level of 10% and above, this may lead to unacceptable quality. The same amount of degradation might be introduced by a conference phone, also adding some noisiness and coloration, and leading once again to unacceptable quality. The other two tested sending devices (headset and mobile phone) are much better, although they also add coloration. The E-model predicts the clean and the G.711 and GSM-EFR-coded conditions to be of good quality; 5% packet loss is predicted to be already below the acceptable threshold (R = 50 in ITU-T Rec. G.109), and the higher packet loss rates as well.

Human speaker identification is affected slightly by GSM-EFR and AMR-NB coding. The figures for Study 2 (which were expected to be intermediate between the figures for the words and the sentences) degrade significantly for packet loss, but also for the conference phone interface, due to discontinuity. They are slightly better for the mobile phone interface and much better for the headset. Automatic speaker recognition degrades considerably with coding; for the AMR-NB, the EER is more than double that of the clean NB channel. Automatic speech recognition is a little less affected, but also degrades with low-bandwidth coding.

C. Analysis for Wideband Channels

When moving to wideband, overall quality considerably improves, as it was expected for this case. The improvement is slightly reduced by coding and by handsets as sending interfaces, but is still above that of good-quality NB transmission. The improvement is predicted both by the POLQA and the E-model, whereas WB-PESQ estimates the clean WB channel too pessimistically. The improved quality is mainly due to the diminished coloration, as DIAL indicates. As in the NB case, the quality degrades considerably with packet loss, and when the conference phone is used as the sending device. DIAL indicates that this is mainly due to an increased discontinuity, but also coloration and noisiness suffer somewhat. POLQA predicts that already connections with 5% packet loss or above would be unacceptable, whereas the E-model considers them still to be acceptable. The conference phone is also predicted to be unacceptable by POLQA.

TABLE 2: SERVICE QUALITY CLASSES AND CORRESPONDING QUALITY AND PERFORMANCE CHARACTERISTICS.

R range	Description	Quality	Human speaker identification	Automatic Recogn.	Examples
>130	Clean SWB	excellent overall quality, no coloration	excellent	unknown	SWB channel
125-130	Clean WB	excellent-good overall quality, no coloration	excellent	excellent	WB channel, G.722.2(23.05)
110-125	WB or SWB coding	good overall quality, low coloration	good-excellent	good	AMR-WB (12.65), G.722 (64), G.729.1
85-95	Toll NB	good overall quality, moderate coloration	good	fair	NB channel, G.711 coding, handset
75-85	NB low-bitrate coding	fair overall quality, stronger coloration	fair-good	fair	AMR-NB coding
50-60	NB interrupted	limiting overall quality, strong discontinuity	fair	unknown	Packet loss < 5%, signal processing
50-75	WB interrupted	fair overall quality, strong discontinuity	good-fair	unknown	Packet loss < 10%, signal processing
<50	Not acceptable	low overall quality, strong discontinuity	fair	unknown	Packet loss > 10%, strong processing artefacts

Human speaker identification significantly improves in the WB case compared to NB. Also when a high rate of packet loss is present, the speaker seems to be far better identifiable than from NB speech. The same applies to the automatic speaker and speech recognition error rates; both are less than half of those observed for NB speech. This shows that WB speech transmission has a clear advantage over NB, not only from an overall quality point-of-view, but also when the speaker has to be recognized (via human or machine), or when the content has to be extracted (automatically). It is expected that a similar improvement can also be observed for human speech intelligibility.

D. *Analysis for Super-wideband Channels*

When moving from WB to SWB, both POLQA and DIAL predict further improvement. However, this improvement may easily be taken away by low-bitrate coding. Unfortunately, only two SWB codecs were tested in Study 2. Human speaker identification scores for the coded versions show that at least this aspect can also be further improved compared to the WB channel. However, more data – also with different codecs, user interfaces and packet loss rates – are necessary to substantiate this finding.

IV. CLASSIFICATION ON A UNIVERSAL SCALE

On the basis of the results of the last section, it becomes possible to classify the transmission conditions with respect to the different criteria. This classification provides guidance as to the improvements which can be achieved for different applications. As transmission planners rely on the transmission rating scale of the E-model which was explained in Section I, the classification of Table 2 is proposed on that basis, starting with the service quality classes defined in ITU-T Rec. G.108 [28] as a starting point, and extending it by the criteria defined in Section II.

According to the table, there are 7 "acceptable" service quality classes ranging from NB to SWB channels. Compared to the "toll NB" case, the WB and SWB channels all offer a higher overall quality, coinciding with lower coloration, better speaker identification, and better automatic recognition of speaker and speech. Below "toll NB", there is the NB coding class characterized by stronger coloration artefacts, as well as two (NB and WB) interrupted classes, characterized by strong discontinuity stemming either from packet-loss or signal-processing degradations. If these degradations become too strong, the overall quality will be unacceptable (R < 50).

V. APPLICATIONS AND FUTURE WORK

The classes defined in Table 2 are helpful for transmission planners. On the basis of the transmission rating ranges and instrumental models like POLQA or the E-model, planners can get an impression which perceptual characteristics can be experienced by prospective users, and which application scenarios the respective channels might be used for. For human telephone conversations, the classes might indicate which codec should be selected, and which packet loss rate is still acceptable. For spoken dialog systems and speaker verification systems, feedback and verification strategies might be selected according to the expected error rates.

Admittedly, there are still a number of gaps in Table 1, which should be filled with more dedicated data. In particular, we think that more codecs and user interfaces, potentially with integrated signal-processing equipment, should be covered. This concerns especially SWB channels for which only very limited data is available. Independently collected data should be used to verify the conclusions and classification made here. This might lead to the definition of a so-called "universal quality scale", with transformation laws to the scores achieved by different prediction algorithms, such as POLQA, WB-PESQ, and the E-model, but also future algorithms for predicting speaker identification or speech intelligibility. Such a universal quality scale has recently been defined as a work item in ITU-T Study Group 12 [29].

In addition, we consider other criteria as relevant for future application scenarios. One very important aspect is of course speech intelligibility, for which relevant assessment methods (and potentially also instrumental prediction models) still need to be defined. Intelligibility is also a major issue for noisy environments, where WB channels can be expected to pick up more noise than NB ones. A further aspect is the extraction of other paralinguistic information about the speaker, such as emotion and personality, both by humans and by automatic recognition algorithms.

REFERENCES

[1] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment Factor Framework for Wideband Speech Codecs", IEEE Trans. Audio, Speech, and Language Proc., 14(6), 1969–1976, 2006.

[2] A. Raake, “Speech Quality of VoIP — Assessment and Prediction”, John Wiley & Sons, Chichester, West Sussex, 2006.

[3] M. Wältermann, I. Tucker, A. Raake, and S. Möller, “Extension of the E-model towards Super-Wideband Speech Transmission”, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP ’10), 4654–4657, Dallas TX, 2010.

[4] “Qualinet White Paper on Definitions of Quality of Experience”, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), P. Le Callet, S. Möller, and A. Perkis, eds., Lausanne, Version 1.2, Novi Sad, March, 2013.

[5] M. Wältermann, “Dimension-based Quality Modeling of Transmitted Speech”, Springer, Berlin, 2013.

[6] M. Wältermann, A. Raake, and S. Möller, “Quality Dimensions of Narrowband and Wideband Speech Transmission”, Acta Acustica united with Acustica, 96(6):1090-1103, 2010.

[7] D. Sen, “Determining the Dimensions of Speech Quality from PCA and MDS Analysis of the Diagnostic Acceptability Measure”, in: Proc. MESAQUIN 2001, Prague, 2001.

[8] ITU-T Rec. G.107, “The E-Model, a Computational Model for Use in Transmission Planning”, International Telecommunication Union, Geneva, 2011.

[9] ITU-T Rec. G.109, “Definition of Categories of Speech Transmission Quality”, International Telecommunication Union, Geneva, 1999.

[10] ITU-T Rec. G.107.1, “Wideband E-Model“, International Telecommunication Union, Geneva, 2011.

[11] M. Wältermann, A. Raake, S. Möller, “Extension of the E-Model Towards Super-Wideband Speech Transmission”, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2010), Dallas TX, 14-19 Mar 2010.

[12] ITU-T Rec. P.835, “Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm”, International Telecommunication Union, Geneva, 2003.

[13] A. Leman, J. Faure, and E. Parizet, “A Non-intrusive Signal-based Model for Speech Quality Evaluation Using Automatic Classification of Background Noises”, In: Proc. 10th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech 2009), 1139-1142, 2009.

[14] H.J.M. Steeneken, and T. Houtgast, “A physical method for measuring speech-transmission quality”, J. Acoust. Soc. Am 67, 318-326, 1980.

[15] IEC 60268-16, “Sound System Equipment – Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index”. International Electrotechnical Commission, Geneva, fourth edition, 2011.

[16] N.R. French, and J.C. Steinberg, “Factors Governing the Intelligibility of Speech Sounds,” J. Acoust. Soc. Am. 19(1):90-119, 1947.

[17] H. Fletcher, R.H. Galt, “The Perception of Speech and its Relation to Telephony“, J. Acoust. Soc. Am 22(2):89-151, 1950.

[18] J. Rodman, “The Effect of Bandwidth on Speech Intelligibility”, White Paper, Polycom, 2006.

[19] ITU-T Contr. COM 12-7, “Benchmark Procedure Proposal for the Assessment of Objective Speech Intelligibility Assessment Methods”, Source: TNO (Author: J. B. Beerends), ITU-T SG12 Meeting, Geneva, 19 – 28 Mar., Geneva, 2013.

[20] ITU-T Rec. P.863, “Perceptual Objective Listening Quality Assessment”, Int. Telecomm. Union, Geneva, 2011.

[21] ITU-T Rec. P.862.2, “Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs”, Int. Telecomm. Union, Geneva, 2007.

[22] N. Côté, “Integral and Diagnostic Intrusive Prediction of Speech Quality”, Springer, Berlin, 2011.

[23] ITU-T Rec. G.191, “Software Tools for Speech and Audio Coding Standardization”, Int. Telecomm. Union, Geneva, 2010.

[24] L. Fernández Gallardo, S. Möller, and M. Wagner, “Comparison of Human Speaker Identification of Known Voices Transmitted Through Narrowband and Wideband Communication Systems”, in: 10. Fachtagung Sprachkommunikation, Braunschweig, ITG im VDE, VDE-Verlag, Berlin, 219-222, 2012.

[25] L. Fernández Gallardo, S. Möller, and M. Wagner, “Human Speaker Identification of Known Voices Transmitted Through Different User Interfaces and Transmission Channels”, in: Proc. 2013 IEEE Int. Conf. Acoust. Speech and Signal Processing (ICASSP 2013), CA-Vancouver, 26-31 May, 2013.

[26] L. Fernández Gallardo, M. Wagner, and S. Möller, “I-Vector Speaker Verification for Speech Degraded by Narrowband and Wideband Channels”, accepted for: ITG Conference on Speech Communication, Erlangen, 2014.

[27] A.V. Ramana, L. Parayitam, and M.S. Pala, “Investigation of Automatic Speech Recognition Performance and Mean Opinion Scores for Different Standard Speech and Audio Codecs”, IETE J Res 58:121-129, 2012.

[28] ITU-T Rec. G.108, “Application of the E-model: A Planning Guide”, International Telecommunication Union, Geneva.

[29] ITU-T Contr. COM 12-41, “Comparability of Quality Indices on the MOS and the R Scale”, Source Deutsche Telekom AG (Author: S. Möller), ITU-T SG12 Meeting, 19 – 28 Mar., Geneva, 2013.