Advances in Perceptual Modeling of Speech Quality in Telecommunications

Hans-Wilhelm Gierlich[*], *Ulrich Heute*[+], *Sebastian Möller*[#]

[*] HEAD acoustics GmbH, 52134 Herzogenrath, Germany
[+] DSS, Technische Fakultät, Christian-Albrechts-Universität zu Kiel, 24143 Kiel, Germany
[#] Quality and Usability Lab, Telekom Innovations Laboratories, TU Berlin, 10587 Berlin, Germany
Email: h.w.gierlich@head-acoustics.de; uh@tf.uni-kiel.de; sebastian.moeller@telekom.de
Web: www.head-acoustics.de; www.dss.tf.uni-kiel.de; www.qu.tu-berlin.de

Abstract

This paper gives an overview about the state of the development of various perception-based quality models. The overview addresses dimension-based modeling and models which specifically focus on quality evaluation in the presence of background noise and modeling echo impairments. Finally the quality evaluation of Text-To-Speech (TTS) systems is discussed.

1 Introduction

Over the last years perception-based models gained increased importance. This is mainly due to advanced signal processing used in telecommunication terminals which does not easily allow the characterization of such devices based on conventional testing. The present paper gives an overview of such models and their state of development.

2 Dimension-Based Modeling of Speech Quality and Technical Cause Analysis

Speech quality can be considered as an event in a multidimensional perceptual space. So, researchers have been interested since a long time in analyzing and describing the dimensions of this space in order to understand the perceptions leading to the quality event (dimension-based modeling), but also to identify the underlying technical causes in order to improve systems so that the perceived quality is optimum (technical cause analysis) [1].

When analyzing modern narrowband and wideband telephone circuits, Wältermann et al. [2] identified three perceptual dimensions which were termed *coloration*, *discontinuity*, and *noisiness*. A fourth dimension related to (non-optimum) *loudness* was added later [3]. Other researchers [4] came up with 6-8 dimensions, separating coloration into high-frequency-absent and low-frequency-absent, and discontinuity into slowly-varying and rapidly-varying fluctuations. Côté [5] developed intrusive estimators for the four dimensions mentioned above, which rely on the input and output signals of the transmission channel to-be-evaluated. He also showed that the estimation of overall quality can be improved by integrating perceptual-dimension estimators into an overall quality estimate. However, although being better that then long-standing standard WB-PESQ [6], his model called DIAL (diagnostic instrumental assessment of listening quality) was outperformed by POLQA [7], the new standard recommended by the International Telecommunication Union, ITU-T. Whereas also POLQA internally calculates 5 parameters (frequency distortions, noise, reverberation distortions, level indicator, spectral flatness indicator – obviously closely related to the above four dimensions) and integrates them with metrics for general distortions, these parameters are not meant to cover the entire space of perceptual quality dimensions as in DIAL. Thus, in addition to POLQA, the ITU-T has started work on a separate Recommendation P.AMD, still in its definition phase, which should propose estimators for perceptual dimensions [8]. The dimension-based approach was also followed by Wältermann [3] who defined a parametric model for network planning, similar to the well-known E-model, however based on three of the four perceptual dimensions defined above.

Whereas perceptual dimensions are helpful to explain perceptual degradations, they may not necessarily uncover the technical causes leading to these degradations. Scholz [9] tried to close this gap by an introduction of "idealized systems", each influencing only one dimension and, by its parameters, directly linking perception and technical effects. Meanwhile, the ITU-T has defined classes of technical causes which can be determined through expert listening [10] . These classes are expected to serve as a basis for automatic technical cause identifiers, to be defined in a future Recommendation P.TCA. Recent analyses [11] have shown that, although there are some relations between perceptual dimensions and technical causes, the moderate values of the correlations indicate that substantial additional information can be extracted when extracting both, perceptual dimensions as well as technical causes. Thus, ITU-T Study Group 12 will follow both parts separately, but perhaps on the basis of common modules, e.g. for signal pre-processing and time alignment, in order to improve the practical applicability of the developed models.

3 Speech Quality in the Presence of Background Noise

One of the most dominant impairments in modern, especially mobile speech communication is noisiness, more specifically, background noise. First contributions on this topic are found already in 2001 in ITU-T [19]. In 2003 an auditory test method was standardized in ITU-T [22] specifically allowing the auditory investigation of speech quality with transmitted background noise (background noise picked up by a free ear at the receiving side was not part of the test method). Auditory tests based on this Recommendation form the basis for the objective prediction of speech quality with transmitted background noise and residual noise-cancellation impairments. The methodology is focusing on speech quality (S-MOS), the intrusiveness of the background noise (N-MOS), and the overall quality (G-MOS). One of the first models trying to objectively predict these quality dimensions was introduced in 2005 [20]. Two bigger projects, aiming at the necessary databases for model development and the development of an objective model allowing an accurate prediction of the S-, N- and G-MOS of these databases, were conducted from 2004 to 2007. As a result, a set of ETSI standards were developed which allow the defined reproduction of background noises in a laboratory environment, include databases with subjective results and a model predicting the S-, N-, and G-MOS values with high correlation to subjective scores [23]. The model is based on the "Relative Approach" [21], a hearing model that is specifically targeted to evaluate perceptually time-variant impairments in the time and frequency domain noticeable by the human auditory system and perceived to be annoying. This hearing model is used for the calculation of S-MOS as well as for N-MOS. A detailed description of the model was published in 2011 [24]. Since the problem became more urgent in modern communication scenarios, esp. when mobile terminals including sophisticated noise cancellation algorithms are used, these ETSI standards also became the basis for terminal-performance evaluation in 3GPP and GSMA. In 2012, a major effort was made collecting new databases for the most recent types of mobile terminals with numerous conditions judged subjectively. These databases formed the basis of an improved objective model still based on the principles described in [24]. This model is targeted specifically to mobile terminals and is standardized in ETSI as well [25]. In all models (also known as 3QUEST), separate model parts exist for the prediction of N-MOS and S-MOS. The G-MOS is calculated using a combination of S- and N-MOS. The major improvement in [25] is the decoupling of the S-MOS calculation from the N-MOS calculation (in the previous models there was always a side-branch modeling the S-MOS partially based on the N-MOS results) and the use of a neural network instead of regression of the model parameters. All models can be applied for a variety of real background noises. Narrowband and wideband communication is covered. Just recently, [25] has been validated further for Music and a single voice as background noise and for Chinese language.

Currently, there is ongoing standardization work in ITU-T to develop a new ITU-T Rec. P.ONRA. This work is targeted to further enhance the applicability of an objective model to super-wideband and additional conditions not yet covered by [23] and [25].

4 Perception-Based Echo-Impairment Prediction

Echo is a talker-related, simultaneous type of impairment and as such perceptually different from other impairments discussed in this paper so far. Echo is a well-known problem, and requirements for echo attenuation exist since many years. However, so far requirements are simply based on attenuation-type measurements (e.g. [26]). It is well known that such pure attenuation-based requirements – not taking into account the masking effect of the own voice – are insufficient and may lead to too demanding requirements preventing more sophisticated signal processing in terminals. Advanced echo control, e.g., might control the echo attenuation by simultaneously taking into account such masking effects. One of the first publications on this topic is found in [29]. PSQM [27] was used as the underlying perceptual model; side-tone and masking were considered basically. A more advanced model is proposed in [24]. The underlying auditory tests for this model are talking and listening tests as well as third-party listening tests as defined in [28]. This model is targeted to the evaluation of terminals. It takes into account the masking effects of the own voice at the talker-side terminal (acoustical and electrical masking) as well as the potential delay of the connection. The model is using the hearing model "Relative Approach" [21] for the masking signal as well as for the echo signal and calculates perceptual differences based on the difference between the outputs of these hearing models. Furthermore, the absolute echo attenuation is used as an input parameter to this model. Ongoing work aims at a further model validation.

5 Quality Evaluation of TTS Systems

In a recent effort, the idea of quality decomposition into dimensions has been transferred from transmitted to synthesized speech. There is an important difference between the two applications: In transmission, the comparison of a "clean" and a "distorted" version of the same signals may serve as the assessment basis. This seems to be feasible also for TTS, if identical utterances from the system and from the same speaker who "fed" the synthesis are available (which usually is not the case). But even then, the synthetic speech may deviate strongly from the natural counterpart, e.g., in emotional or prosodic details, without a deteriorated listening quality. So, "single-ended" or "non-intrusive" methods are needed.

Such techniques have also been developed for speech-transmission quality [14]. Here, a large set of parameters is measured acoustically from the single speech signal, some "key parameters" lead to a pre-classification of the system at hand, and a combination of other values allows to predict the listening-only quality "objectively". This strategy is also applicable to TTS signals; especially, the parameters describing a vocal-tract model, evaluated for their "natural" behavior, were shown to be useful [15].

This procedure may also be combined with the fundamental attribute approach. In [16], this was investigated, and details of a thorough study of various approaches towards a dimension-based quality assessment are shown. The basis is given by three (main) dimensions found from auditory evaluation, multidimensional-scaling (MDS), and semantic-differential experiments at TU Berlin [17]. These dimensions were termed *naturalness of voice*, absence of *disturbances*, and *calmness*, according to the correlations between – a priori abstract – dimensions and a list of attributes (see Tab. 1). More and differing attributes with, however, strong relations to the above ones, were found dependent on the type of TTS systems and signals included in the tests (see Tab. 2).

Attribute	Dimension 1		Dimension 2		Dimension 3	
scales	female	male	female	male	Female	male
acceleration	0.54	0.74	0.76	0.83	-0.53	
bumpiness	-0.61		-0.81	-0.79		
clink	-0.67	-0.60	-0.56		0.62	
distortions	-0.72	-0.77	-0.77	-0.70		
disturbances	-0.63	-0.69	-0.77	-0.69		
fluency	0.59	0.53	0.88	0.78	-0.53	
hiss	-0.54					
intelligibility	0.83	0.87	0.73	0.64		
naturalness	0.85	0.85	0.77	0.88		
noise						
pleasantness	0.58	0.87	0.78	0.78		
polyphony	-0.64	-0.63	-0.77	-0.58	0.54	
rasping	-0.56	-0.63				
rhythm	0.80	0.76	0.86	0.84	-0.55	
speeed					0.54	0.65
tension	-0.71	-0.51	-0.59	-0.59	0.68	0.54
overall impression	0.90	0.89	0.76	0.76		

Tab. 1: Correlations between 3 dimensions of an MDS and a list of given attributes [18].

	Dimension 1	**Dimension 2**	**Dimension 3**	**Dimension 4**	**Dimension 5**
Tag	**Naturalness of voice**	**Prosodic quality**	**Fluency and intelligibility**	**Absence of disturbances**	**Calmness**
Relevant Scales	Naturalness, voice pleasantness	Accentuation, rhythm, pros., intonation	Fluency, intellig., bumpiness, polyphon.	Hissing, noise, rasp. disturbances	Speed, Tension
Test I	Naturalness		Temporal distortions	Disturbances	Speed
Test II	Naturalness of voice	Temporal distortions			Calmness
Test III	Naturalness &	prosody	Intelligibility		
Audio-book	Listening pleasure	Prosody & rhythm			

Tab. 2: Different dimensions as found from the analysis of different test material [16].

On this basis, various approaches towards a construction of dimensional and integral quality estimators and their evaluation by cross-validations are discussed. The best versions achieve a correlation of up to 0.96 with subjective assessments. These most promising approaches apply a support-vector machine on a non-linear combination of non-linearly transformed signal measurands. The latter non-linearity stems from the novel concept of "perceptual regularization" – essentially a 1-bit quantization of features perceived as "non-regular" (0) or "regular" (1). The most useful features turned out to be MFCCs and some prosodic terms [16] .

6 Conclusions and Future Work

With respect to the estimation of perceptual dimensions and technical causes, a number of issues need to be resolved, namely:

- In the current requirement specification of P.AMD [8], there are two sets of perceptual-dimension estimators, with two corresponding subjective evaluation methods. As the effort in conducting such subjective experiments which could serve as a ground truth for the estimators to be developed is high, it remains to be seen how reliable the basis for this Recommendation will be. The fact that ITU-T SG12 has opted for a collaborative approach can be seen as an advantage in this respect.
- As P.AMD is considered as an add-on to POLQA, its dimensions estimators are expected to rely on the same signals, i.e. the input and the output signals of the channel under consideration. For practical applications, it would be desirable to complement such a model with non-intrusive estimators which rely only on the degraded output signal.
- For P.TCA, there are still a number of dimensions which cannot yet be extracted with good reliability [12]. For the technical causes where inter-rater agreement is low, exemplary speech files would be an advantage. Such files have been produced for 31 of the 47 classes of technical causes defined in P.TCA [12]. For the other causes, corresponding material is still missing.
- Both P.AMD and P.TCA are limited to degradations which can be dealt with in a purely listening-only situation. The underlying reason is that the dimensions [2] as well as the technical causes [10] have been identified in such a situation. For a talking or interacting situation, respective subjective methods which could identify underlying dimensions are still missing or in practical validation [24]. Approaches presented in [13] still await their psychophysical validation.
- The ETSI models [23] and [25] are widely used for terminal evaluation in different areas. The new work currently ongoing in ITU-T to further enhance the models and expand their application might lead to an even advanced method allowing to predict speech quality in the presence of background noise.
- The prediction of echo impairments based on a perceptual model as described in [24] just started. The model is applied for terminals especially in the area

of research and algorithm development. Currently no standardization activity exists on this topic.

- The feasibility of an instrumental perceptual model also for TTS signals was shown in [16]. The coarse regularization applied up to now, however, as well as the robustness of the estimators need further work.
- Only very few models explicitly model cognitive effects. Work on modelling such effects is currently underway in the ABBA project [30].

References

[1] U. Heute, S. Möller, A. Raake, K. Scholz, M. Wältermann, "Integral and Diagnostic Speech-Quality Measurement : State of the Art, Problems, and New Approaches", in: *Proc. Forum Acusticum 2005*, Budapest, pp. 1695-1700, 2005.

[2] M. Wältermann, A. Raake, S. Möller, "Quality Dimensions of Narrowband and Wideband Speech Transmission", *Acta Acustica united with Acustica*, 96(6), 1090–1103, 2010.

[3] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*, Springer, Berlin, 2013.

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

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

[6] 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.

[7] ITU-T Rec. P.863, *Perceptual Objective Listening Quality Assessment (POLQA)*, Int. Telecomm. Union, Geneva, 2011.

[8] ITU-T COM 12-274, *Draft Requirement Specification for P.AMD (Perceptual Approaches for Multi-Dimensional Analysis)*, Source: Deutsche Telekom AG (S. Möller, M. Wältermann), Int. Telecomm. Union, Geneva, 2011.

[9] Scholz, K., "Instrumental Quality Assessment of Telephone-Band Speech based on Quality Attributes", Dissertation (in German), CAU Kiel, Shaker Publ., Aachen, 2008.

[10] ITU-T TD 650rev1 (GEN/12), *Requirement Specifications for P.TCA (Technical Cause Analysis)*, Source: Rapporteur Q.16/12 (L. Malfait), Int. Telecomm. Union, Geneva, 2011.

[11] S. Möller, F. Köster, F. Schiffner, J. Skowronek, "Analyzing Technical Causes and Perceptual Dimensions for Diagnosing the Quality of Transmitted Speech", in: *Proc. 4th Int. Workshop on Perceptual Quality of Systems (PQS 2013)*, Wien, 2-4 Sept, pp. 30-35, 2013.

[12] ITU-T COM 12-105, "P.TCA Exemplary Listening Material Processing and Validation", Source: Deutsche Telekom AG (F. Köster, F. Schiffner, J. Skowronek, S. Möller), Int. Telecomm. Union, Geneva, 2013.

[13] Köster, F., Möller, S., "Towards a New Test Paradigm for the Subjective Quality Assessment of Conversational Speech", in: *Fortschritte der Akustik – DAGA 2013: Plenarvortr. u. Fachbeitr. d. 39. Dtsch. Jahrestg. f. Akust.*, Meran, Dtsch. Ges. Akust., Berlin, 2013.

[14] ITU-T Rec. P.563, *Single-Ended Method for Objective Speech-Quality Assessment in Narrow-Band Telephony Applications*, Int. Telecomm. Union, Geneva, 2011.

[15] Norrenbrock, C., Heute, U., Hinterleitner, F., *On the Use of Vocal-Tract Approximations for Instrumental Quality Assessment*, ITG Conf. Speech Comm., Braunschweig, 2012.

[16] Norrenbrock, C., *Instrumental Quality Estimation for Synthesized Speech Signals*, Dissertation, CAU Kiel, Shaker Publ., Aachen, 2013.

[17] Hinterleitner, F., Möller, S., Norrenbrock, C., Heute, U., *Perceptual Quality Dimensions of Text-to-Speech Systems,* Proc. Interspeech, pp. 2177-2180, 2011.

[18] Hinterleitner, F., Norrenbrock, C., Möller, S., Heute, U., *Multidimensional Analysis of Perceptual Quality of Text-to-Speech Systems,* Proc. Interspeech, pp. ???, 2012.

[19] Gierlich, H.W., *Evaluation of the Quality of Background Noise Transmission using the "Relative Approach"*, ITU-T COM 12-34-E, October 2001.

[20] Gautier-Turbin, V., Le Faucheur, N., *A Perceptual Objective Measure for Noise Reduction systems*, Mesaqin 2005, Prague.

[21] Genuit, K.: "*Objective Evaluation of Acoustic Quality Based on a Relative Approach*", InterNoise '96, Liverpool, UK.

[22] ITU-T Rec. P.835, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, Int. Telecomm. Union, Geneva, 2003.

[23] ETSI EG 202 396-1,2,-3 standard series: *Speech Quality in the Presence of Background Noise*, ETSI 2005-2007

[24] Reimes, J., Gierlich, H.W., Kettler, F., Poschen, S. Lepage, M., *The Relative Approach Algorithm and ist Applications in New Perceptual Models for Noisy Speech and Echo Performance*, Acta Acustica united with Acustica, Vol. 97 (2011) pp. 325-341

[25] ETSI TS 103 106, *Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods*, ETSI 2012

[26] ITU-T Rec. G.114, *Talker Echo and Its Control)*, Int. Telecomm. Union, Geneva, 2003.

[27] ITU-T Rec. P. 862, Perceptual evaluation of speech quality (PESQ): *An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, Int. Telecomm. Union, Geneva, 2001.

[28] ITU-T Rec. P. 831, *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*, Int. Telecomm. Union, Geneva, 1998.

[29] EP 1216519 A1, "*Measuring the perceptual quality of speech signals including echo disturbances*", European Patent, 1999.

[30] A. Raake, J. Blauert, "Comprehensive Modelling of the Formation Process of Sound Quality", *Proc. QoMEx 2013*, Klagenfurt, 2013.