Dimension-based Diagnostic Prediction of Speech Quality

Sebastian Möller, Ulrich Heute

Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, 10587 Berlin
DSS, Technische Fakultät, Christian-Albrechts-Universität zu Kiel, 24143 Kiel
Email: sebastian.moeller@telekom.de, uh@tf.uni-kiel.de
Web: www.qu.tu-berlin.de, www.dss.tf.uni-kiel.de

Abstract

In this paper, we address models for predicting the quality of speech transmission and communication services. In contrast to models which aim at predicting the *integral quality* of speech, i.e., at an index for the overall quality, we focus on models which predict individual dimensions of quality, as they are subjectively perceived by a communication partner. Such predictions of individual dimensions can then be combined to form an integral- quality estimation; however, they may also be used separately in order to diagnose the reasons of insufficient quality. Beyond, describing particular perceptual effects rather than technical systems, they are less vulnerable by new system variants not included in the model development.

1 Motivation

Starting point of our approach (see Heute et al. [1]) is the observation that standard speech-quality prediction algorithms provide little insight into the reasons for a specific quality judgment. In addition, the development of quality-prediction models during the past 30 years shows that always new algorithms had to be developed because the existing ones did not cover perceptual effects which were provoked by new coding, transmission, or, more general, processing equipment, such as medium bit-rate codecs, packet loss, or noise-reduction algorithms.

We are thus seeking algorithms which are *robust* enough to generalize towards new types of impairments, and *diagnostic* enough to uncover the reasons of bad quality. For this aim, two approaches are conceivable:

- Algorithms identifying *technical* system characteristics which result in impairments, such as bandwidth restrictions, noise levels, interruptions, etc.
- Algorithms identifying *perceptual* dimensions related to impairments, such as degraded sound color, noisiness, or interruptions of the conversational flow.

It is not obvious to decide whether the technical system characteristics or the perceptual dimensions better satisfy the requirements of robustness and diagnostic power as outlined above. An analysis of the perceptual dimensions extracted from transmitted speech over the past 40 years (see e.g. [1]) shows that – apart from the effect of interruptions due to discontinuous transmission – perceptual dimensions seem to be stable in the long run. Signal-processing algorithms and their impact on the transmitted signals, on the other hand, change very quickly. We are thus convinced that algorithms which are based on perceptual dimensions rather than technical characteristics will provide a better generalizability in the long run, and thus better satisfy the *robustness* requirement. In turn, they might not be overly diagnostic, in the sense that they would be able to determine the root of an impairment down to the device or algorithm that caused that particular impairment. Such information, however, may again be retrieved, since the dimension-impairment descriptions themselves are necessarily based on technical impacts.

2 Identification of Dimensions

2.1 Transmitted Natural Speech Signals

Perceptual dimensions from transmitted or processed natural speech have been frequently identified in the past, following one of two approaches:
(1) By scaling perceptual differences of pairwise-presented stimuli, and then mapping the perceptual distance to a multidimensional space using the multidimensional-scaling procedure (MDS, see [3]); or
(2) by asking listeners to rate all stimuli independently on a set of bipolar scales (so-called Semantic Differential, SD) and reducing the space of judgments with the help of a factor analysis and subsequent rotation.
Whereas the former procedure always results in orthogonal dimensions, the latter does not always do so, depending on the rotation used.

Applying both of these procedures to both narrowband and wideband-transmitted speech stimuli, Wältermann et al. [4] identified three dimensions representative for speech signals transmitted through state-of-the-art channels, namely *coloration*, *noisiness*, and *discontinuity*. A fourth dimension was added later representing the *loudness* of the transmitted signals. Sen [5] used the Diagnostic Acceptability Measure (DAM, a variation of the SD technique) developed by Voiers [6], and identified between 4 and 7 dimensions; in particular, he subdivided coloration into ‘low frequencies absent’ and ‘high frequencies absent’, continuity into ‘slowly-varying characteristics’ and ‘rapidly-varying characteristics’, and noisiness into ‘background noise’ and ‘signal-correlated noise’. Depending on which level of detail one seeks for, one or the other set of dimensions might be more relevant.

In practice, measuring dimension values for a larger number of stimuli is hardly feasible with the MDS and SD techniques, because of the high experimental effort involved. As an alternative, Wältermann showed that the four dimensions of their set can also be scaled directly by listeners, with high reliability [7]. The technique is similar to what is currently recommended for noisy speech signals in ITU-T Rec. P.835 [8], where listeners are asked to rate the quality of the speech signal, of the background noise, and of the overall stimulus in a sequence, on three separate scales.

Identifying dimensions becomes more complicated when considering speech transmission channels in actual usage situations, i.e. bidirectional conversations. A simple assumption is that the listening-quality dimensions

discussed before add up with dimensions related to the talking-only quality (resulting e.g. from impaired sidetone or echo) and dimensions related to the sequence of talkspurts in a conversation. However, this cannot easily be demonstrated, as there are no subjective methods which would allow identifying dimensions in a talking or conversational situation. As a consequence, we cannot assume that the dimensions are orthogonal (in fact, talking-only degradations will almost certainly influence also the conversational behavior).

2.2. Synthetic Speech Signals

In recent experiments, the dimension-based approach has also been extended to speech synthesized with Text-To-Speech (TTS) systems. Here, the focus is not on the effect of the transmission channel, but on the speech-production process itself. In a neutralized setting, Hinterleitner et al. [9] identified *naturalness*, *temporal degradations* and noise-like *disturbances* as the (non-orthogonal) dimensions. Whereas there seems to be a link to the three dimensions from Wältermann et al. mentioned above, the first dimension *naturalness* is obviously not an orthogonal description of just one effect. It will mostly be influenced by the prosody of the synthesized signal, which is not the case for naturally-produced speech (where the prosody is always expected to be "natural"). It will also depend on the TTS-system's ability to avoid "unnatural-voice" effects, like hoarseness, or breathiness, which, on the other hand, are partly covered by *temporal degradations*.

In a more application-like setting of an audiobook-reading task, Hinterleitner et al. [10] identified the dimensions *listening comfort* ("Hörgenuss") and *prosody* as the most influential ones. This shows that a dimension-based analysis may only be relevant if the dimensions are chosen according to their (expected) relevance for the application under consideration. Different names, though, do not mean a strong contradiction: Listening comfortably needs certainly a natural voice and low distortions.

3 Prediction of Dimensions

A variety of approaches have been presented in the past which aim at predicting individual dimensions. According to the required input information, they can be differentiated into

- full-reference signal-based models, relying on the input and output signal of the channel or, more generally, of the processing,
- no-reference signal-based models relying on only the degraded (output) signal, and
- parametric approaches which rely on parametric descriptions of the elements of the channel or the processing system.

3.1 Natural Speech Signals

For the dimensions extracted by Wältermann et al., Scholz [11] described initial estimators which capture the perceptual effects at least for some technological devices which could have produced them (e.g., zero-insertion packet-loss concealment in case of discontinuity, filters of general shape for coloration, etc.). These estimators have been improved by Côté [2] and combined with an estimator for the integral quality in order to provide reliable estimations both for the integral quality and its underlying perceptual dimensions. The model proved to be better than the then-known signal-based standard Wideband-PESQ (ITU-T Rec. P.862.2) on a large number of databases, but was still outperformed by the new standard POLQA (ITU-T Rec. P.863) which focusses on the integral quality only. This shows that a dimension-based approach may come at some (limited) costs in comparison to a non-dimension-based one, but with an increased diagnostic value. This may, however, be overcome, if the contributing measurements are used for both the dimension estimations and, independently, for a direct total-quality estimation, as indicated by Scholz [11]. For noisy speech, the model in [12] which is based on the so-called "Relative Approach" [13], provides a good performance in estimating the three dimensions addressed in ITU-T Rec. P.835.

On the parametric side, Wältermann [7] showed that the three dimensions cited above may also be reliably estimated with parameters which are commonly used in network planning. Using parametric estimations of the three dimensions, he developed a dimension-based version of the E-model, a parametric tool for transmission planning [14]. The combination of dimensions towards integral quality was performed using a Euclidean norm of a positive vector describing the respective degradation of each dimension. The results on a limited set of databases showed that the dimension-based approach could even outperform the E-model; still, the approach needs to be validated on a larger set of independent test data.

When it comes to conversational aspects, Guéguin [15] developed a model which combines listening-only, talking-only, and conversational dimensions by translating them to the "transmission rating scale" underlying the E-model, and then summing up the respective dimensions. Still, the model did not show a sufficient performance in case of background noise, which required additional steps (noise detection and bypassing of the talking-quality dimension estimate) to be taken. The model described in [13] also shows how echo degradations can be predicted instrumentally. However, no dimension-based model for the entire conversational situation is yet available which would show a sufficiently high performance.

3.2 Synthetic Speech Signals

For natural signals, the use of "single-ended", i.e., no-reference quality measurements has been tried [16], though with little success, compared to the reference-based methods named above: As long as a reference with "normal" sound is available, the comparison of the "clean" and "distorted" signals usually outperforms all trials to directly derive defects from a deteriorated signal. For TTS signals, however, there is, in many cases, no "clean" prototype available.

The idea has been investigated by Seget [17] to use a natural speaker of "similarly" sounding voice as a reference, with limited success. Even if, as in rare applications, the original speaker's version of the same sentences is available, the results are not yet convincing [18]. Fur-

ther work towards a refined time-structure alignment might help.

First estimations of individual dimensions related to TTS quality have been provided, using general parameters known from speech-signal analysis, modulation spectra, and voice-quality related parameters [19]. While the results derived from the envelope spectrum show a high dependence on the datasets used for training and testing, voice analysis is a good candidate for the naturalness description. Apparently, the task is far more difficult than the quality prediction for transmitted natural speech.

4 Discussion and New Questions

4.1 Present Quality Estimators

In Fig. 1, the well-known performance of two signal-comparison based estimators is shown, which do not use any dimension analysis. Fig. 2 shows that a similar, though reduced performance is possible with a dimension basis, which, however, opens additional diagnostic insight. Fig. 3, finally, demonstrates the possibility to measure one of the essential dimensions instrumentally also for synthetic speech, without use of any reference.

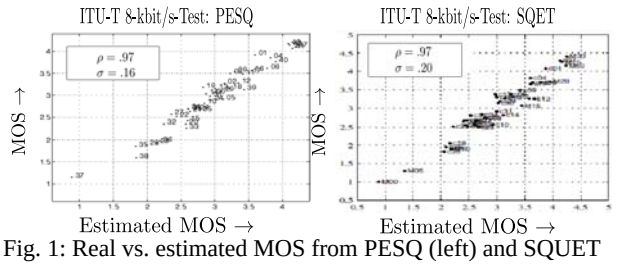

Fig. 1: Real vs. estimated MOS from PESQ (left) and SQUET (right; Hauenstein [20]).

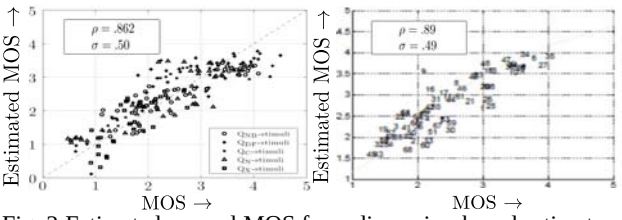

Fig. 2:Estimated vs. real MOS from dimension-based estimators based on coloration, noisiness, and discontinuity: Work by Scholz (left, [11]) and Côté (right, [2]).

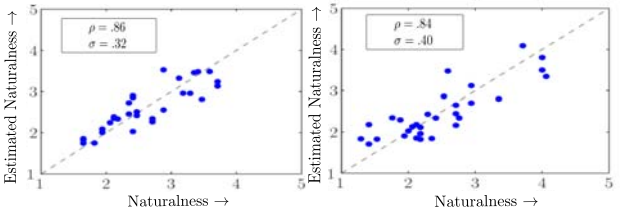

Fig. 3: Estimation of the dimension “naturalness” for TTS systems: female (left) and male (right) voices (Norrenbrock et al. [21]).

4.2 Subdimensions

One possible way towards further improvements is an investigation of the dimensions themselves. As stated above, they are not necessarily orthogonal, and they are obviously not (always) representing separable effects. So, they may be decomposed into “sub-dimensions”. For example, “coloration” may consist of band limitations, low- or high-frequency emphasis, or fluctuations inside a certain band, giving rise to effects like “dullness”, “sharpness”, or “distance”. Such ideas were already followed (e.g., by Scholz [11]), also for “noisiness” (Kühnel [22]), and they are deepened in recent work on all three above attributes and on narrow-band as well as wide-band speech signals (Huo [23]). For synthetic speech, the inclusion of “voice-quality” descriptions or of “many relevant” parameters from ITU-T Rec. P.563 [16] follows the same line [24].

The question whether these “sub”-dimensions are in fact the “true” (orthogonal?) dimensions, may sound academic, but is worth some work – if the necessary huge data bases are available and the MDS or SD experiments are affordable (or simplified).

4.3 Robustness of the Approach

In the introduction, we raised the question whether a prediction approach based on perceptual dimensions would be more robust than one which is based on technical causes. One indication for this might be the extensibility of approaches when moving from narrowband (300-3400 Hz) to wideband (50-7000 Hz) speech transmission. The standard PESQ model has early been adapted by simply replacing the input filter with a wideband version and modifying the mapping function; however, the validity of the extended model was limited so that it was felt necessary to set up a completely new standard, ITU-T Rec. P.863 (POLQA).

For the dimension-based approach from Wältermann et al. [4], the dimensions extracted in narrowband tests were shown to be mainly valid also for the wideband case. Thus, the DIAL model developed by Côté [2] relies on the same perceptual dimensions for both narrowband and wideband speech. In fact, the model was shown to outperform PESQ better in the wideband than in the narrowband case. Starting from a theoretical discussion about the perceptual quality space, Wältermann [7] shows that moving from narrowband to wideband in equivalent to a translation along the “coloration” dimension. He expects that a further shift along this dimension would also make the perceptual E-model fit for super-wideband speech.

Whereas the dimensions might be stable on the long run, the integration function which provides a weighting of the dimensions with respect to their impact on overall quality might need adjustments when the transmission technologies – and with it the experiences of the listeners – develop. Such an adaptation might be easy when explicit weights are provided, e.g. in the linear approach selected by Scholz [11]. In turn, when machine-learning techniques are used (as it was the case in Côté [2]), a retraining will require a large set of databases which reflect the new equipment and the experiences of the listeners.

4.4 Current Developments

Dimension-based approaches are currently experiencing a boost in the world of standardization. Three such work items are currently under study in ITU-T Study Group 12. The first item (P.ONRA) tries at estimating speech quality and noise quality, and combining them to an estimate for noisy speech as it would be judged according to the P.835 paradigm. The second item (P.AMD) tries to set up a set of estimators for the 4…7 dimensions of transmitted

speech given in Section 2. The third item (P.TCA) tries to identify technical causes of degradations. Both P.AMD and P.TCA are expected to cover the whole range of degradations which is currently taken into account by POLQA in ITU-T Rec. P.863.

When extending the approach to the conversational situation, new subjective paradigms are necessary in order to come up with meaningful dimensions. For this purpose, an analytic method is needed which nonetheless reflects the conversational situation (including talking and change of talkspurts) in a valid way. Further research is necessary to transfer the approach also to audio-visual speech services, e.g. video conferencing or video telephony. We expect that it is not enough to simply add dimensions which have been identified individually for the auditory and visual signals. For example, the synchrony between audio and video, as well as the importance of audio and video for the aim of the conversation, are important aspects which should be taken into account by a proper weighting of the dimension-integration model.

Acknowledgment

The work behind this overview is sponsored in parts by "Deutsche Forschungsgemeinschaft", with DFG grants Mo 1038/2-1, 5-2, 11-1, He-4465 / 1-1, 1-2, 4-1). This is gratefully acknowledged.

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, HU-Budapest, 1695-1700.

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

[3] J. Carroll, J., *Individual Differences and Multidimensional Scaling.* Multidimensional Scaling —Theory and Applications in the Behavioral Sciences Volume I —Theory (R.N. Shepard, A.K. Romney, S.B. Nerlove, eds.), 105–155, 1972.

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

[5] D. Sen, *Determining the Dimensions of Speech Quality from PCA and MDS Analysis of the Diagnostic Acceptability Measure*. In: Proc. MESAQUIN 2001, CZ-Prague.

[6] W.D. Voiers, *Diagnostic Acceptability Measure for Speech Communication Systems.* In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP) 1977, 204–207, US-Hartfort CT.

[7] M. Wältermann, *Dimension-based Quality Modelling of Transmitted Speech*. Doctoral dissertation, TU Berlin, DE-Berlin, 2012.

[8] ITU-T Rec. 835, *Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm*, Int. Telecomm. Union, CH-Geneva, 2003.

[9] F. Hinterleitner, S. Möller, C. Norrenbrock, U. Heute, *Perceptual Quality Dimensions of Text-to-Speech Systems*. In: Proc. Interspeech 2011, IT-Florence, 2177-2180.

[10] F. Hinterleitner, G. Neitzel, S. Möller, C. Norrenbrock, *An Evaluation Protocol for the Subjective Assessment of Text-to-Speech in Audiobook Reading Tasks*. In: Proc. Blizzard Workshop 2011, IT-Torino, 2011b.

[11] K. Scholz, *Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen*. Doctoral dissertation, CAU Kiel (Arbeiten über Digitale Signalverarbeitung Nr. 32, U. Heute, ed.), Shaker Verlag, DE-Aachen, 2008.

[12] EG 202 396-3, *Speech Quality Performance in the Presence of Background Noise. Part 3: Background Noise Transmission - Objective Test Methods*. Europ. Telecomm. Standards Inst., FR-Sophia Antipolis, 2008.

[13] J. Reimes, H.-W. Gierlich, F. Kettler, S. Poschen, M. Lepage, *The Relative Approach Algorithm and its Applications in New Perceptual Models for Noisy Speech and Echo Performance*. Acta Acustica united with Acustica, 97(2), 325-341, 2011.

[14] ITU-T Rec. G.107, *The E-Model, a Computational Model for Use in Transmission Planning*, Int. Telecomm. Union, CH-Geneva, 2005.

[15] M. Guéguin, *Évaluation objective de la qualité vocale en contexte de conversation*. Doctoral dissertation, Université de Rennes 1, FR-Rennes, 2006.

[16] ITU-T Rec. P.563, *Single-Ended Method for Objective Speech-Quality Assessment,* Int. Telecomm. Union, CH-Geneva, 2004.

[17] K. Seget, *Untersuchungen zur auditiven Qualität von Sprachsyntheseverfahren,* Dipl. Thesis, CAU Kiel, DE-Kiel, 2007.

[18] F. Hinterleitner, S. Zabel, S. Möller, L. Leutelt, C. Norrenbrock, *Predicting the Quality of Synthesized Speech Using Reference-Based Prediction Measures.* In: Proc. ESSV 2011, DE-Aachen, 99-106.

[19] C. Norrenbrock, U. Heute, F. Hinterleitner, S. Möller, *Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals.* IEEE Signal Processing Letters, accepted for publication, 2012.

[20] M. Hauenstein, *Psychoakustisch motivierte Maße zur instrumentellen Sprachgütebeurteilung,* Doctoral dissertation, CAU Kiel (Arb. ü. Digitale Signalverarbeitung Nr. 10, U. Heute, ed.), Shaker Verlag, DE-Aachen, 1997.

[21] C. Norrenbrock, U. Heute, F. Hinterleitner, S. Möller, *Aperiodicity Analysisfor Quality Estimation of Text-to-Speech Signals.* In : Proc. Interspeech 2011, IT-Florence, 2193-2196.

[22] C. Kühnel, K. Scholz, U. Heute, *Dimension-based Speech Quality Assessment: Development of an Estimator of the Dimension "Noisiness"*. In: Proc. ITG Conf. Speech Comm. 2008, DE-Aachen.

[23] L. Huo, *Attribute-based Speech-Quality Assessment – Narrowband and Wideband.* Doctoral dissertation, CAU Kiel (Arbeiten über Digitale Signalverarbeitung Nr. 37, U. Heute, ed.), Shaker Verlag, DE-Aachen, 2012.

[24] C. Norrenbrock, *On the Use of Vocal Tract Approximations for Instrumental Quality Assessment.* Accepted for publication in: Proc. ITG Conf. Speech Comm. 2012, DE-Braunschweig.