Non-Intrusive Estimation Model for the Speech-Quality Dimension Loudness

Friedemann Köster, Victor Cercos-Llombart, Gabriel Mittag, Sebastian Möller

Quality and Usability Lab, Technische Universität Berlin, 10587 Berlin
Email: {friedemann.koester,victor.cercos,gabriel.mittag,moeller}@tu-berlin.de
Web: www.qu.tu-berlin.de

Abstract

In this article, we present an approach towards a new non-intrusive speech quality estimator. The proposed method facilitates the evaluation of speech telephony services and provides diagnostic information by assessing dimensions of the perceptual quality space. One of these quality dimensions is *Loudness*, which describes a non optimal sound level. As an important part of the proposed model, a non-intrusive *Loudness* estimator is presented. The estimator uses a linear regression with five different indicators that are extracted from the output signal only, to map subjective *Loudness* judgments. The new model is trained on one and tested on two independent subjective databases. In addition, the performance of the *Loudness* estimator is compared to the diagnostic intrusive quality estimator *Diagnostic Intrusive Assessment of Listening quality* (DIAL). The evaluation shows that the estimator provides results on a high reliability level, indicating the applicability and the value of the proposed estimator for diagnostic enhancement.

1 Introduction

The quality of transmitted speech is a critical parameter for telecommunication system providers to evaluate their services for costumer satisfaction, reputation, and economical reasons. Thus, it is important to assess the quality of transmitted speech in vocal human-to-human telephony communication as perceived by the end-user, the so-called *Quality of Experience* (QoE) [1]. Traditionally, QoE feedback is gathered in subjective laboratory experiments [2]. Participants are asked to listen to stimuli that describe the system under test and to rate the perceived overall quality. The most common procedures are based on the *Absolute Category Rating* (ACR) task and lead to the *Mean Opinion Score* (MOS), representing the average rating for an average person for each signal or processing condition [3]. While subjective experiments are reliable means to gather system users feedback they demand significant effort in terms of time and money. Hence, instrumental signal-based methods have been established. They are divided into two groups depending on the input signal they require [4]:

- *Intrusive* methods, also known as double-ended methods, since they use a reference signal (clean or system input) and a corresponding degraded signal (distorted or system output). They compare the signals and map the differences to a predicted MOS rating.
- *Non-intrusive* methods, also known as single-ended methods, since they only use the degraded signal. They map specific signal characteristics of the signal to a predicted MOS rating.

For intrusive models, the *International Telecommunication Union* (ITU-T) recommended the long-term standard *Perceptual Estimation of Speech Quality* (PESQ) for *narrowband* (NB, 300-3,400 Hz) and WB-PESQ for *wideband* (WB, 50-7,000 Hz) transmission channels [5, 6]. It is now replaced by its successor *Perceptual Objective Listening Quality Assessment* (POLQA) that also considers *super-wide-band* (SWB, 50-14,000 Hz) transmissions [7]. However, intrusive models exhibit certain inherent limitations for practice [8]:

- If the resulting MOS estimation is suboptimal it provides little insight into the reason for the quality-loss. More precisely, two different degraded speech stimuli (one degraded by e.g. a low bit rate codec and the other by e.g. noise) can both be rated with the same (low) MOS value. No diagnostic information is provided.
- The intrusive models demand that the non-degraded input signal of the transmission channel is available making them only valuable in a laboratory context. For practical online monitoring applications only non-intrusive models that do not require the input signal are useful.

In this paper we present an approach that overcomes both limitations. We introduce a non-intrusive speech quality estimator based on perceptual dimensions for diagnosing transmitted speech. After giving a short overview of related work and an introduction into perceptual dimensions, the new approach is introduced. As a subsequent and important part of the proposed model a non-intrusive estimator for the perceptual dimension *Loudness* and its results will be presented. The results are discussed and conclusions regarding the non-intrusive overall speech quality estimation are drawn.

2 Related Work

The aforementioned limitations are not new to the community and have been subject to various research. However, they were mostly only considered separately. On the one hand, it was investigated how to provide diagnostic information, on the other hand non-intrusive methods to estimate the overall quality have been developed.

One idea to provide diagnostic information is to analyze the *perceptual quality space* of transmitted speech. It results from several studies that speech quality is by nature a multidimensional space. This multidimensional space is formed by a number of orthogonal *quality features*, called *perceptual dimensions* [9]. The underlying assumption is that the perceptual dimensions are directly related to their impairing *quality elements*, like codecs, filters, packet loss, or noise. Assessing the perceptual dimensions thus serves as diagnosis [10].

However, the main problem met in the definition of the perceptual quality space is the characterization and identification of the perceptual dimensions. Two methods are conceivable: (I) gathering similarity or preference measures and conducting a subsequent *multidimensional scaling* [11], or (II) conducting a *principal component analysis* on ratings stemming from a *semantic differential* experiment [12]. In [13] both methods were applied to NB and WB transmission systems and three perceptual dimensions

ISBN 978-3-8007-4275-2

were identified:

- *Noisiness*: describes degradation such as background noise, circuit noise or coding noise. It is labeled with "not noisy" and "noisy".
- *Discontinuity*: describes degradations concerning isolated or non-stationary distortions introduced by e.g. the loss of packets. It is labeled with "continuous" and "discontinuous".
- *Coloration*: describes degradation resulting from frequency response distortions, introduced by e.g. bandwidth limitations. It is labeled with "uncolored" and "colored".

However, in [13] the stimuli used in the experiments were presented at a normalized listening level. In several other studies [14–16] the listening level is also considered as a quality feature of the overall quality of transmitted speech. Thus, the perceptual dimension *Loudness* was included to the perceptual space defined by three aforementioned dimensions. However, this perceptual dimension can be correlated with other perceptual dimensions, its orthogonality is not proven.

- *Loudness*: is important for the overall quality and the intelligibility. It is labeled with "optimum sound level" and "non-optimum sound level".

These four dimensions should reflect the whole perceptual quality space used by a subject in order to rate the quality of transmitted speech signal [4]. The four dimensions are also defined in the current ITU-T work item *Perceptual Approaches for Multi-Dimensional Analysis* (P.AMD) that targets at standardizing a diagnostic intrusive method [17]. The introduced quality profile allows to map the overall quality on the basis of perceptual dimensions giving the possibility to identify reasons for quality losses. In addition, [10] created a subjective test-paradigms to directly quantify the perceptual dimensions. Yielding from these subjective experiments the intrusive model *Diagnostic Instrumental Assessment of Listening quality* (DIAL) that predicts the overall quality as well as the introduced perceptual dimensions has been developed [4]. It serves as reference for the proposed *Loudness* estimator.

To provide a standardized non-intrusive overall quality model the ITU-T performed a competition in 2004 that produced two submissions. One is the now recommended standard ITU-T P.563 [18] that generates an internal reference as replacement for the missing input signal using LPC-analysis and showed to be reliable for NB telecommunication scenarios. The other competitor is called *Auditory Non-Intrusive QUality Estimation* (ANIQUE) and uses the approach of modeling the representation of the speech signal at the central level of the human auditory system with the temporal envelope representation of the speech signal [19]. Both methods are only recommended for NB transmission systems and estimate only the overall quality MOS, no diagnostic information are provided. To provide a non-intrusive model that is also suitable for WB and SWB speech transmission ITU-T currently launched a new standardization process [20].

The estimation of the *Loudness* has been part of several studies in the past. In [21] the *Loudness* is estimated for stationary sounds, in [22] a method to estimate the *Loudness* in time varying sounds is developed, and in [23] an algorithm to estimate the *Loudness* in audio samples is standardized, to name just a few. In the this work we adopt some of these approaches and optimize them to estimate the *Loudness* in the whole transmitted speech signal.

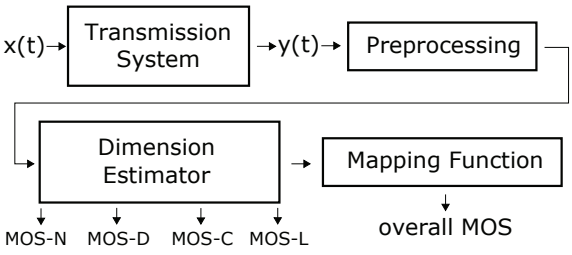

Figure 1: *Structure of the new non-intrusive estimator. N-Noisiness, D-Discontinuity, C-Coloration, L-Loudness.*

3 Non-intrusive Speech Quality Estimator

The far-end objective of this work is to overcome the aforementioned limitations. For this, a new approach of non-intrusive quality prediction is planned. The scope of the new estimator will cover three main points: (a) providing diagnostic information, (b) considering WB and SWB transmission channels, as well as (c) giving the possibility of practical monitoring operations. A draft overview of the planned model structure can be seen in Figure 1. The model consists of three fundamental blocks:

- Preprocessing: filtering and level alignments as well as separation of active and non-active segments via *Voice Activity Detection* (VAD).
- Dimension estimator: four sub-blocks, each block is one estimator for one perceptual dimension. D- *Discontinuity*, N- *Noisiness*, L- *Loudness*, C- *Coloration*.
- Mapping function: the individual estimations of each perceptual dimensions will be mapped to the overall speech quality using their coherency described in [10].

The output of the new model will consist of five values, one for each dimension and the overall MOS. To realize the proposed model, individual dimension estimators for each perceptual dimension have to be developed first. Since a non-intrusive model is planned, we follow the approach of identifying interpretable indicators (signal characteristics) of the output speech signal that can be mapped to the corresponding perceptual dimension. This concept was applied in two previous studies that presented the development of estimators for the dimensions *Coloration* [24] and *Noisiness* [25]. The dimension estimators showed that the concept provides reliable results on unknown and independent data. As the next fundamental step the development for a dimensions estimator for *Loudness* is presented in the next sections.

4 Evaluation Data

The proposed *Loudness* estimator is evaluated on three databases. The databases were part of the POLQA competition and served for training and testing [4]. They consist of swiss-german speech samples with multiple sentences (double sentences, duration: 8 s - 12 s) and four speakers. The subjective overall quality MOS is rated according to [3] and the four subjective dimension MOS values are gathered following the proposed test-paradigm presented in [10]. The evaluation is thus conducted on subjective *Loudness* ratings on double-sentence and condition base. The three databases *Swiss01*, *Swiss02* and *Swiss03* are mixed-band (NB, WB, and SWB) and contain signal-

ISBN 978-3-8007-4275-2

Database	Conditions	Stimuli	Speaker	Hours of Speech
	Training-Data			
$Swiss01$	50	200	4	≈ 0.44 h
	Test-Data			
$Swiss02$	50	200	4	≈ 0.44 h
$Swiss03$	54	216	4	≈ 0.48 h

Table 1: Overview of the five databases.

correlated, uncorrelated as well as ambient background noise, temporal clipping, different coding, packet loss, different frequency distortions, and combinations of these degradations. The covered degradations were chosen to meet the scope of the POLQA model [7]. We decided to mix the data and used one databases (*Swiss01*) for training and two databases (*Swiss02*, and *Swiss03*) for testing. An overview of the databases can be seen in Table 1.

5 Indicators

As mentioned before, first interpretable indicators that describe the *Loudness* – non optimal sound level, like stationary or time varying level fluctuations – have to be determined. These indicators must be extracted from the output speech signal only. After dividing the speech signal in active and inactive segments with the VAD the indicators that best fit a linear regression model to map the *Loudness* have to be found. For this, a *repeated sequential cross-validation feature selection* is applied. The method splits the training data (*Swiss01*) into one third training and two thirds test-sets (k-cross-validation $K = 3$). Subsequently, the method sequentially selects indicators ($x \in P$) to best fit a linear regression model:

$$\widehat{MOS_L} = \hat{f}_m(P) = \beta_0 + \sum_{j=1}^{J} \beta_j x_j \tag{1}$$

As long as the *Pearson* correlation ρ [26] between the estimated *Loudness* MOS values $\widehat{MOS_L}$ and the subjective *Loudness* MOS values MOS_L is increasing, the method adds parameter to the regression model:

$$\rho(MOS_L, \widehat{MOS_L}) = \frac{1}{K} \sum_{k=1}^{K} \rho(MOS_L^{(k)}, \widehat{MOS_L^{(k)}}) \tag{2}$$

If the correlation reaches a determined threshold, the algorithm stops. The procedure is repeated 100 times to vary training and test sets (cross-validation). This is done to be certain that a variety of combinations of the samples is considered. The method outputs the average importance of each indicator. For the *Loudness* the methods identified five indicators:

MAdB (Maximum Amplitude in Decibel): It is assumed that the maximum value gives an assumption about a non-optimal sound level. Thus, this indicator describes the maximum value (peak) of all samples ($k \in K$) in one speech signal. Equation 3 shows how the $MAdb$ indicator is calculated.

$$MAdB = 10 \cdot log_{10}(max_K(x(k))) \tag{3}$$

MAL (Mean Active Level): To capture level variations, the whole speech signal has to be analyzed. However, only speech active segments should be considered, because the calculation of the speech level in inactive segments distorts the calculation. Therefore, for this indicator the arithmetic mean absolute amplitude in all active speech segments ($k_A \in K_A$) is calculated:

$$MAL = \frac{1}{K_A} \sum_{i=1}^{K_A} |x(i)| \tag{4}$$

ARMS (Active Root Mean Square): The root mean square, also called quadratic mean, is defined as the square root of the arithmetic mean. In comparison to the MAL indicator, higher amplitudes are considered with a higher weight. Hence, the $ARMS$ indicator is also supposed to capture the sound level in the whole speech signal. Again, only active speech segments are considered:

$$ARMS = \sqrt{\frac{1}{K_A} \sum_{i=1}^{K_A} |x(i)|^2} \tag{5}$$

MALB (Mean Active Level on the Bark Scale): The relation between frequency and place on the basilar membrane is neither linear nor logarithmic. For describing different phenomena of psychoacoustics, it has proven to be very useful to apply a transformation of frequency onto a scale showing a linear relation to the place of basilar membrane excitation [27]. One of the most widely used examples of such a scale is the *Bark scale* [21]. To capture this effect for the *Loudness*, active speech segments are transformed in the frequency domain and than expressed on the Bark scale. The arithmetic mean transformed in decibel values gives the MALB indicator.

SPL (Sound Pressure Level): The sound pressure level is a logarithmic measure of the RMS of a sound relative to a reference value ($\hat{p}_0 = 2 \cdot 10^{-5}$ Pa) [21]. For the description of the auditory sensitivity one can plot the sound pressure level that is perceived at a certain loudness, over frequency. Equation 6 shows how the SPL indicator is calculated.

$$SPL = 20 \cdot log_{10}(\frac{RMS}{\hat{p}_0}) \tag{6}$$

6 Estimator

The feature selection algorithm presented above allows to identify the five introduced indicators. In addition, the method provides the regression coefficients (β_0 and β_js in Equation 1) for a linear regression model. The model is trained on the training database to best map the subjective ratings for the perceptual dimension *Loudness*. Based on the results of the selection algorithm we developed a *Loudness* estimator using the five introduced parameter and the corresponding regression coefficients (each making a significant contribution, $p < 0.05$):

$$\begin{aligned} \widehat{MOS}_{Lou} = &-7.25 - 0.06 \cdot \textbf{MAdB} \\ &+ 10.12 \cdot \textbf{MAL} - 37.15 \cdot \textbf{ARMS} \\ &+ 0.04 \cdot \textbf{MALB} + 0.18 \cdot \textbf{SPL} \end{aligned} \tag{7}$$

To evaluate the estimator and analyze its performance the *Loudness* estimator (Equation 7) is then applied on the two test databases *Swiss02* and *Swiss03*. Additionally, the

ISBN 978-3-8007-4275-2

Table 2: *Results of the Loudness estimator and the DIAL model.*

Database	Proposed Estimator Correlation ρ	RMSE	DIAL Correlation ρ	RMSE
$Swiss02$	0.92	0.31	0.91	0.70
$Swiss03$	0.93	0.32	0.93	0.55
all	0.92	0.32	0.92	0.62

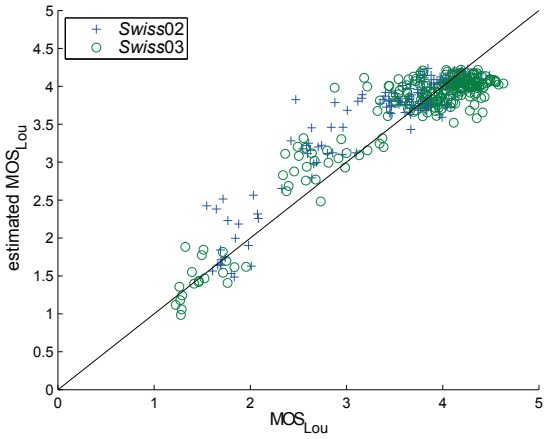

Figure 2: *Results of the Loudness estimator.*

data is merged and the performance is also evaluated on both databases jointly. The results can be seen in Table 2 and Figure 2. To give a reference of the performance we calculated the *Loudness* with the intrusive DIAL model. The DIAL *Loudness* estimations can also be seen in Table 2.

The resulting estimated $\widehat{MOS_{Lou}}$ values show high correlations with the subjective MOS_{Lou} values with a total correlation of 0.92 and Root Mean Square Errors (RMSE)s below 0.32 for all databases. On both databases solid correlation values above 0.9 were archived. This shows that the introduced indicators applied in a linear regression seem to robustly map the *Loudness*.

The proposed model reaches about the same correlations as the DIAL model. Looking at the RMSEs though, the new *Loudness* model outperforms the DAIL model. One explanation for that could be, that the DIAL model uses only two indicators to predict the *Loudness*. Having more indicators at hand seems to provide more precise estimations.

Looking at Figure 2 it can be seen that most speech samples of the two test databases are located in the high rated MOS regions between 4 and 4.5. These ratings are very well covered by the proposed estimator. For lower rated *Loudness* values it seems as if the estimator is slightly data dependent. While the samples of the *Swiss03* database are covered well (green circles), the estimator is slightly over-predicting the samples of the *Swiss02* database (blue crosses). This can especially be seen for estimations of samples with subjective MOS_{Lou} values around 2 and 3.

Overall, respecting the different data and the non-intrusive approach, the results show that the proposed estimator is capable of predicting the perceptual dimension *Loudness*.

7 Conclusions

The objective of the work presented in this article is to develop a new non-intrusive speech quality estimator. The scope of the proposed estimator is to provide diagnostic information and to give online monitoring possibilities. Thus, a non-intrusive estimator based on perceptual dimensions is introduced. While first estimators for the perceptual dimensions *Coloration* and *Noisiness* have already been developed, as a subsequent step a *Loudness* estimator and its corresponding indicators are introduced. The results of the evaluation show that the *Loudness* estimator provides correlations and errors on a high reliability level. Considering the complications of a non-intrusive approach and the number of used data the proposed estimator is a major landmark towards a overall non-intrusive quality estimator.

Together with the estimators for *Coloration* and *Noisiness* we now have three out of four dimension estimators at hand. As the next final step a non-intrusive *Discontinuity* estimator is demanded. Apart from that, we will also introduce non-linear regression approaches that might reduce the error and increase the correlation. Having all four estimators and a final regression method at hand we strive to map the overall quality resulting in a non-intrusive speech quality estimator.

8 Acknowledgments

The presented work was supported by the Federal Ministry of Education and Research, Germany (01IS12056) and the Software Campus.

References

[1] S. Möller and A. Raake, *Quality of Experience: Advanced Concepts, Applications and Methods*. Berlin: Springer, 2014.

[2] S. Möller, *Quality Engineering*. Berlin: Springer, 2010.

[3] ITU-T Recommandation P.800, *Methods for Subjective Determination of Transmission Quality*. Geneva: International Telecommunication Union, 1996.

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

[5] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*. Geneva: International Telecommunication Union, 2001.

[6] ITU-T Recommendation P.862.2, *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. Geneva: International Telecommunication Union, 2007.

[7] ITU-T Recommendation P.863, *Perceptual Objective Listening Quality Assessment*. Geneva: International Telecommunication Union, 2011.

[8] F. Köster, S. Möller, J.-N. Antons, S. Arndt, D. Guse, and B. Weiss, "Methods for Assessing the Quality of Transmitted Speech and of Speech Communication Services," *Acoustics Australia*, vol. 42, pp. 179 – 184, December 2014.

[9] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. Berlin: Springer Science & Business Media, 2005.

[10] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*. Berlin: Springer, 2012.

[11] I. Borg and P. Groenen, *Modern Multidimensional Scaling - Theory and Applications*. New York, NY: Springer Series in Statistics, 2005.

[12] C. Osgood, *The Measurement of Meaning*. Urbana, IL: University of Illinois Press, 1957.

[13] M. Wältermann, A. Raake, and S. Möller, "Quality Dimensions of Narrowband and Wideband Speech Transmission," pp. 1090–1103, Acta Acustica united with Acustica, 2010.
[14] B. McDermott, "Multidimensional analyses of circuit quality judgements," *JASA*, vol. 3, no. 45, pp. 774–781, 1996.
[15] N. Côté, V. Gautier-Turbin, and S. Möller, "Influence of loudness level on overall quality of transmitted speech," in *Proc. 123rd AES Convention*, (USA-New York,NY), 2007.
[16] E. H. Rothauser, G. E. Urbanek, and W. P. Pachl, "Isopreference method for speech evaluation," *Journal of the Acoustical Society of America*, vol. 44, no. 2, pp. 408 – 418, 1968.
[17] ITU-T Temporary Document TD 438rev1 (GEN/12), *Requirement Specifications for P.AMD (Perceptual Approaches for Multi-Dimensional Analysis)*. International Telecommunication Union; Rapporteur Q.9/12 (J. Berger), 2014.
[18] ITU-T Recommendation P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*. Geneva: International Telecommunication Union, 2004.
[19] D. S. Kim and A. Tarraf, "Perceptual model for non-intrusive speech quality assessment," (Montreal, Canada), Proc. IEEE ICASSP, 2004.
[20] ITU-T Temporary Document TD (GEN/14), *Technical Requirement Specification P.SPELQ*. International Telecommunication Union; Rapporteur Q.9/12 (J. Berger), 2014.
[21] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Springer-Verlag Berlin Heidelberg, 3 ed., 2007.
[22] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *J. Audio Eng. Soc*, vol. 50, no. 5, pp. 331–342, 2002.
[23] ITU-R Recommandation BS.1770-1, *Algorithms to measure audio programme loudness and true-peak audio level*. Geneva: International Telecommunication Union, 2007.
[24] G. Mittag, F. Köster, and S. Möller, "Non-intrusive Estimation of the Perceptual Dimension Coloration," in *Fortschritte der Akustik, DAGA 2016: Plenarvortr. u. Fachbeitr. d. 42. Dtsch. Jahrestg. f. Akust.*, DEGA, 2016.
[25] F. Köster, G. Mittag, T. Polzehl, and S. Möller, *Non-intrusive Estimation of Noisiness as a Perceptual Quality Dimension of Transmitted Speech*. Proc. 5th International Workshop on Perceptual Quality of Systems (PQS 2016), submitted, 2016.
[26] J. Bortz, *Statistik: für Human- und Sozialwissenschaftler (Springer-Lehrbuch) (German Edition)*. Springer, 6., vollst. überarb. u. aktualisierte aufl. ed., 9 2004.
[27] A. Raake, *Speech Quality of VoIP Assessment and Prediction*. Chichister, West Sussex: John Wiley & Sons, 2006.

ISBN 978-3-8007-4275-2