Non-intrusive Estimation of the Perceptual Dimension Coloration

Gabriel Mittag, Friedemann Köster, Sebastian Möller
Quality and Usability Lab, Technische Universität Berlin, Deutschland,
Email: gabriel.mittag@gmail.com, friedemann.koester@tu-berlin.de, sebastian.moeller@tu-berlin.de

Abstract

In this article, we present a new method for the non-intrusive quality estimation of transmitted speech. The proposed method provides diagnostic information and facilitates the evaluation of speech telephony services. For diagnostic information the approach of assessing perceptual quality-relevant dimensions is used. One of these quality dimensions is the Coloration that describes degradations resulting from frequency response distortions, like bandwidth limitations. As part of the new method, a non-intrusive parametric Coloration estimator is presented. The estimator is trained on two and tested on three independent subjective databases. Additionally, the performance of the estimator is compared to the diagnostic intrusive quality estimator DIAL (Diagnostic Intrusive Assessment of Listening quality). The results show that the estimator provides a high reliability level, indicating the applicability and the value of the proposed estimator for diagnostic enhancement.

Introduction

Inside a telephone service (human vocal-to-vocal communication over landline, mobile, or *VoIP* connections), the quality of transmitted speech can be impaired by different network and terminal devices. These *quality elements* [1] are for example codecs, bandwidth limitations, linear and non-linear filters, packet loss, noise, and others [2]. In this context, the quality of transmitted speech as perceived by the service users, the so-called *Quality of Experience* (QoE) [3], is an important parameter for telephone service providers to evaluate their systems.
Traditionally, QoE is assessed with subjective *listening-only tests* (LOTs) in a laboratory context. Naïve participants rate the perceived overall quality of a speech sample on a *Abolute Category Rating* scale. The ratings are averaged in the so-called *Mean Opinion Score* (MOS) [4], representing the average rating of an "average" service user.
However, subjective quality assessment methods are money and time consuming. Thus, instrumental models that estimate the outcome of a subjective LOTs have been established. These so-called signal-based models, can be divided into two groups: (I) *Intrusive* models rely on the input and the output signal of a transmission channel. They compare the signals and map the differences to a predicted MOS rating. (II) *Non-intrusive* models rely only on the (degraded) output signal of a transmission channel. They map specific signal characteristics to a predicted MOS rating.
The *International Telecommunication Union* (ITU-T) recommended the long-term standard intrusive model *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. It has now been replaced by its successor *Perceptual Objective Listening Quality Assessment* (POLQA) that additionally covers super-wideband ((SWB) 50-14.000 Hz) channels. In practice however, these intrusive standards exhibit certain inherent limitations:
(1.) The resulting MOS prediction provides few insights into the reason for possible quality-losses.
(2.) Intrusive models demand the availability of the degradation-free input signal. This limits the usage for monitoring purposes as the input signal is often only available in the laboratory.
In this article we present the approach of overcoming both limitations. The basic idea is to establish a non-intrusive speech quality estimator that is based on perceptual dimensions for diagnosing the quality of transmitted speech. First, we give an overview about related work and introduce the concept of perceptual dimensions. Second, the new approach and its fundamental structure is shown. As the first part of the new model, the non-intrusive estimation of the perceptual dimension Coloration and its results are presented. The article closes with a conclusion and an outlook into the future.

Related Work

In the past, the two aforementioned limitations have been part of various research. To overcome the first limitation we investigated how to provide diagnostic information to provide more information about quality-losses. For the second limitation non-intrusive models for estimating the MOS were developed.
For obtaining diagnostic information the approach of assessing quality relevant *perceptual dimensions* was applied. The underlying idea is as follows: Users of a telephone service face a *sound event*. This sound event can be degraded by the aforementioned quality elements and causes a *perceptual event* inside the listener. The perceptual event is of a multidimensional nature and is composed of explicit *perceptual features*. The features are directly connected to specific quality elements and thus serve for diagnosis [1]. If the features are orthogonal they are referred to as perceptual dimensions [5].
To identify perceptual dimensions two experimental paradigms ((a) *Pairwise Similarity* with *Multidimensional Scaling*; and (b) *Semantic Differential* with a *Principal Component Analysis*) were applied to NB and WB transmission channels resulting in four dimensions [5]:

- *Noisiness*: describes degradations like background noise, circuit noise, or coding noise.
- *Discontinuity*: describes degradations concerning isolated or non-stationary distortions, introduced by e.g. the loss of packets.
- *Coloration*: describes degradations resulting from frequency response distortions, introduced by e.g. bandwidth limitations.
- *Loudness*: is important for the overall quality and the intelligibility.

The four dimensions form a new quality profile that allows to also map the overall quality on the basis of the perceptual dimensions. This gives the possibility to diagnose the quality and identify reasons for quality-losses. It was shown in [6] that it is possible to directly quantify the identified dimensions in subjective tests. For this, participants rate each dimension on a individual scale, similar to what is proposed for noisy signals in [7].

Yielded from the results of subjective experiments the intrusive diagnostic instrumental model *Diagnostic Instrumental Assessment of Listening quality* (DIAL) has been developed [8]. The model predicts the overall quality as well as the introduced perceptual dimensions.

Looking at the second limitation, intrusive models are not useful for the online monitoring proposes since the input speech signal of a transmission channel is largely unavailable.

To provide new monitoring models, the ITU-T performed a competition to standardize a non-intrusive method in 2004 that produced two submissions, both are only recommended for NB speech transmission and provide no diagnostic information. One is the now recommended standard ITU-T P.563 [9]. The algorithm generates an internal reference as replacement for the missing input signal using LPC-analysis. The second 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 [10].

Planned Non-Intrusive Speech Quality Estimator

For overcoming the mentioned limitations of traditional estimators, a novel approach of non-intrusive quality estimation is considered. We plan to establish a model that respects diagnostic information, WB and SWB channels, as well as practical monitoring operations. A structure of the proposed model can be seen in Figure 1. The structure consists of three blocks:

- Preprocessing; filtering and level alignments as well as separation of active and non-active segments via *Voice Activity Detection*.
- Dimension estimator: four sub-blocks, each block is one estimator for one perceptual dimension.
- Mapping function: separate estimations of the perceptual dimensions will be mapped to the overall quality MOS using the coherency described in [6].

To provide diagnostic information the model will output one MOS value for each perceptual dimension in addition to the overall MOS. The idea of the four dimension estimators is to identify interpretable parameters (signal characteristics) of the output signal $y(t)$ that can be mapped to the corresponding dimension. As a first step, this concept will be explained for the dimension Coloration in the next sections.

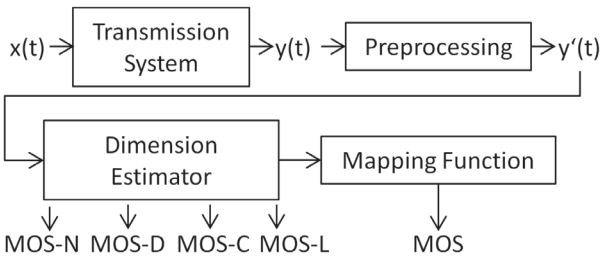

Figure 1: Structure of the new non-intrusive estimator. N-Noisiness, D-Discontinuity, C-Coloration, L-Loudness.

Database	Conditions	Stimuli	Speaker	hours of speech
	Training-Data			
Swiss01	50	200	4	≈ 0.44 h
DAT1	66	792	12	≈ 1.76 h
	Test-Data			
Swiss02	50	200	4	≈ 0.44 h
Swiss03	54	216	4	≈ 0.48 h
DAT2	76	912	12	≈ 2.03 h

Table 1: Overview of the five databases.

Data

Five databases are available to evaluate the Coloration estimator. All databases represent state of the art degradations like different NB, WB and SWB conditions, noise, packet loss or band-pass filter. They consist of processed and live-recorded speech samples with different sentences (double sentences, duration: 8 s - 12 s) and speakers (4 - 12) as well as condition based subjective ratings for the perceptual dimensions gathered following the the paradigm presented in [6].

Two databases (**DAT1** and **DAT2**) were created to identify the perceptual dimensions in [6]. The other three databases (**Swiss01**, **Swiss02** and **Swiss03**) were used to validate the POLQA model [8]. We decided to mix the data and use two databases (DAT1 and Swiss01) for training and three databases (DAT2, Swiss02, and Swiss03) for testing. An overview of the databases can be seen in Table 1.

Coloration Parameter

As a first step, interpretable parameters that describe the Coloration - isolated or non-stationary distortions like bandwidth limitations or signal correlated disturbances - and that can be extracted from the output signal only, have to be determined. Since the extracted parameters alone cannot robustly map the Coloration, an optimal combination must be found. Thus, we applied a *repeated sequential cross-validation feature selection*. This method splits the training data into training- and test-sets (cross-validation) and sequentially selects parameters that best fit a linear regression model. Then, parameters are added

as long as the *Pearson* correlation between the estimated Coloration (result from the regression model) and the subjective Coloration ratings rises. The algorithm stops when the correlation can not be further increased (determined threshold) by adding additional parameters. The method is repeated 100 times with varying training and test sets. Finally, the algorithm indicated the parameters most frequently used and the mean correlation for the training sets. Applying this procedure, 8 parameters that are important for mapping the Coloration were identified:

BW (BandWidth): First, the logarithmic *Power Spectral Density* (PSD) for all inactive speech segments with a length of 32 ms is calculated and transformed onto the bark scale. The cut off frequency is then specified as the point in the PSD where the level falls below half that of the maximum. In the case that the PSD has a significant notch, the search algorithm ignores it to achieve a more robust estimate. The distance in between the cut off frequencies in bark is then extracted as feature BW (BandWidth) (cf. Figure 2).

PED (Pitch Envelope Deviation): This parameter assesses signal-correlated disturbances (that influence the perceived Coloration) in voiced speech segments. It is assumed, that in clean signals the difference between the peaks of the pitch and its smoothed envelope is smaller than in noisy signals. For this, the peaks of the pitch envelope are found and then a cubic smoothing spline is calculated. The median of the normalized absolute error for each peak location is the PED value.

FV (Frequency Variation): Speech signals with signal correlated noises often have a flat noise floor for higher frequencies [11], which also seem to have an impact on the perception of the Coloration. To grasp this effect the absolute sum of difference of the PSD ($P_{yy}(\mu)$) is calculated for frequencies from μ_1=500 Hz to μ_2=4000 Hz. Then the difference between the maximum and minimum is subtracted to obtain a feature that is less dependent on the individual speech signal, see Equation 1.

$$FV = \left(\sum_{\mu=\mu_{1+1}}^{\mu_2} abs(P_{yy}(\mu) - P_{yy}(\mu-1)) \right) - (max(P_{yy}(\mu)) - min(P_{yy}(\mu))) \quad (1)$$

CEP-STDi, **CEP-STDa**, and **CEP-KUTa** (Cepstrum): The cepstrum for each 5 ms frame is calculated according to [12]. Then, the statistical measurements *standard deviation* (STD) and *kurtosis* (KUT) are extracted from each cepstrum. The "per inactive frame" (i) and "per active frame" (a) averages of the statistical measurements are then used as the three features.

LPC-SKEWi (Linear Prediction): An 8th order LPC Model is calculated for each 5 ms frame. Then, similar to the cepstrum features, the statistical measurement *skewness* (SKEW) for inactive (i) frames is extracted.

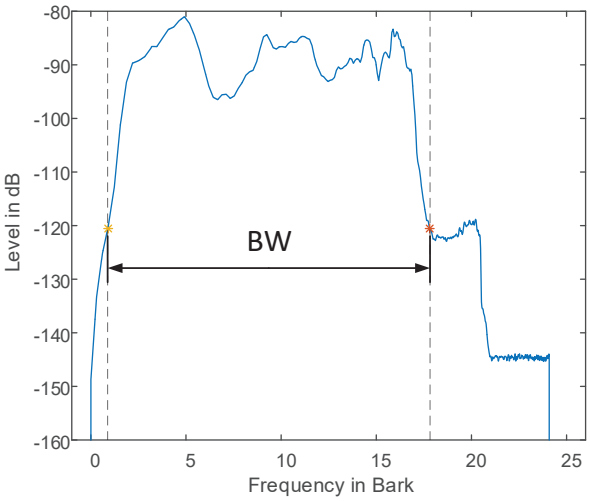

Figure 2: Calculation of the parameter BW (Bandwidth).

MFCC-STDa (Mel Frequency Cepstral Coefficients): This feature is motivated by the work of [13]. MFCC is a perceptual-based speech analysis which is widely used, e.g. for speech and speaker recognition [14]. For the feature, the coefficients of a 5th order MFCC model are calculated for each frame with 32 ms length. Then their standard deviations (without the zeroth coefficient) are calculated and the "per active frame" (a) average is extracted.

Coloration Estimator

Through applying the feature selection algorithm on the training databases it is possible to identify the most important features and to eliminate the ones with a negative influence on the correlation. Additionally, the selection method provides the optimal coefficients for a linear regression model to best map the perceptual dimension Coloration. Based on this data, we applied a Coloration estimator using the 8 introduced parameter and the corresponding regression coefficients (each making a significant contribution, $p < 0.05$):

$$\begin{aligned} \widehat{MOS}_{Col} = {} & 29.63 + 0.18 \cdot \mathbf{LPC\text{-}SKEWi} \\ & - 53.83 \cdot \mathbf{CEP\text{-}STDi} - 1.44 \cdot 10^{-9} \cdot \mathbf{FV}^4 \\ & - 4.71 \cdot \mathbf{CEP\text{-}STDa} + 0.08 \cdot \mathbf{BW} \\ & - 1.29 \cdot \mathbf{MFCC\text{-}STDa} - 53.88 \cdot \mathbf{PED} \\ & + 25.2 \cdot \mathbf{CEP\text{-}STDi}^2 + 246.68 \cdot \mathbf{PED}^2 \\ & + 0.15 \cdot \mathbf{CEP\text{-}KUTa} \end{aligned} \quad (2)$$

The Coloration estimator is then applied on the three test databases separately and jointly. The results can be seen in Table 2 and Figure 3. To compare the results, the intrusive DIAL model was used as a reference. The DIAL Coloration estimations can also be seen in Table 2. The resulting estimated $\widehat{MOS}_{Col}$ values show high correlations with the subjective MOS_{Col} values with a total correlation of 0.93 and RMSEs below 0.40 for all databases. Furthermore, the resulting correlations on all three test databases provide consistent values

Database	Proposed Estimator Correlation ρ	RMSE	DIAL Correlation ρ	RMSE
*Swiss*02	0.90	0.37	0.81	0.85
*Swiss*03	0.92	0.36	0.89	0.35
*DAT*2	0.98	0.40	0.98	0.23
all	0.93	0.38	0.89	0.47

Table 2: Results of the Coloration estimator.

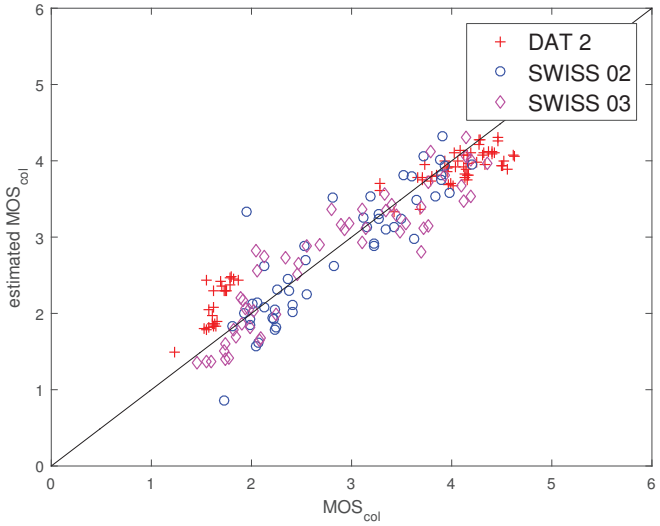

Figure 3: Results of the Coloration estimator.

greater than 0.9. The proposed model outperforms the DIAL model for database Swiss02 and Swiss03 in terms of correlation and *Root Mean Squared Error* (RMSE). For database DAT2 the proposed model achieves the same correlation, but a higher RMSE than the DIAL model. This can be seen in Figure 3 where low Coloration ratings are slightly overestimated (red crosses). Overall, respecting the different data and the non-intrusive approach, the results show that the proposed estimator is capable of predicting the perceptual dimension Coloration.

Conclusion

In this contribution a new approach towards the non-intrusive estimation of speech quality is presented. As a first step a Coloration estimator and its corresponding parameters are introduced. Through applying the proposed method, it is now possible for the first time to estimate subjective ratings of the perceptual dimension Coloration with a non-intrusive approach. The results of the evaluation show that the estimator provides correlations and errors with a high level of reliability. Regarding the complications of a non-intrusive approach on the one hand, as well as the rather high number of databases included on the other hand, the proposed estimator is a major landmark towards a integrated non-intrusive quality estimator.

In future work we plan to further introduce non-linear regression approaches that might reduce the error and increase the correlation. As a next step, estimators for Discontinuity, Noisiness and Loudness are required. Having these predictors at hand we strive to map the overall quality, resulting in a non-intrusive speech quality estimator.

References

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

[2] A. Raake, *Speech Quality of VoIP Assessment and Prediction*, John Wiley & Sons, Chichister, West Sussex, 2006.

[3] Qualinet, "Qualinet White Paper on Definitions of Quality of Experience," 2013, (Version 1.2, eds. P. Le Callet, S. Möller, A. Perkins), European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland.

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

[5] M. Wältermann, A. Raake, and S. Möller, "Quality Dimensions of Narrowband and Wideband Speech Transmission," Acta Acustica united with Acustica, 2010, pp. 1090–1103.

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

[7] ITU-T Recommendation P.835, *Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Supression Algorithm*, International Telecommunication Union, Geneva, 2003.

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

[9] ITU-T Recommendation P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*, International Telecommunication Union, Geneva, 2004.

[10] D. S. Kim and A. Tarraf, "Perceptual model for non-intrusive speech quality assessment," Montreal, Canada, 2004, Proc. IEEE ICASSP.

[11] S. Voran, "Observations on the t-reference condition for speech coder evaluation," *CCITT SG XII Experts Group on Speech Quality, Document Number SQ 13.92*, 1992.

[12] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing (2Nd Ed.)*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.

[13] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods.," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 6, pp. 1935–1947, 2006.

[14] K. Gopalan, T.R. Anderson, and E. Cupples, "A comparison of speaker identification results using features based on cepstrum and fourier-bessel expansion," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 289 – 294, 1999.