INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Non-intrusive Diagnostic Monitoring of Fullband Speech Quality

Sebastian Möller[1,2], *Tobias Hübschen*[3], *Thilo Michael*[1], *Gabriel Mittag*[1], *Gerhard Schmidt*[3]

[1]Quality and Usability Lab, Technische Universität Berlin
[2]Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI)
[3]Digital Signal Processing and System Theory, Christian-Albrechts-Universität zu Kiel
{sebastian.moeller|thilo.michael|gabriel.mittag}@tu-berlin.de
{thu|gus}@tf.uni-kiel.de

Abstract

With the advent of speech communication systems transmitting the full audible frequency band (0-20,000 Hz), traditional approaches for narrowband (300-3,400 Hz) speech quality estimation, service planning and monitoring come to their limits. Recently, signal-based as well as parametric tools have been developed for fullband speech quality prediction. These tools estimate overall quality, but do not provide diagnostic information about the technical causes of degradations. In the present paper, we evaluate approaches for diagnostically monitoring the quality of super-wideband and fullband speech communication services. The aim is, first, to estimate technical causes of degradations from the degraded output signals, and, second, to combine the estimated causes with parametric quality prediction models to obtain a quantitative diagnostic picture of the quality-degrading aspects. We evaluate approaches for non-intrusively identifying coding schemes and packet-loss, and compare estimated quality to the predictions of an intrusive signal-based model.

Index Terms: speech quality prediction, super-wideband, fullband, diagnosis, coding, packet-loss, non-intrusive

1. Introduction

Speech communication services have made use of a restricted narrowband audio channel for more than 100 years. This situation has changed with the advent of packet-based transmission techniques, which are able to not only transmit signals through the standard narrowband (300-3,400 Hz) audio channel, but as well wideband (50-7,000 Hz), super-wideband (20-14,000 Hz) or fullband (0-20,000 Hz) signals. The transmission of this extended bandwidth can generally be expected to improve quality; however, IP-based transmission and the applied coding schemes also come with degradations of the (improved) quality. Thus, the quality of super-wideband and fullband speech communication services remains an important topic, and requires new tools to be evaluated.

Quality engineering of speech communication services can be performed in different phases of service set-up and operation. During early planning phases, only planning values of the equipment which is expected to be used are available, typically in terms of tabulated parameter values, such as attenuations, delay times, codecs used, expected probabilities of packet loss, etc. In this phase, parametric planning models such as the E-model [1] are used to predict the quality of a communication service using this equipment. As soon as simulations of actual services become available (e.g. codecs, channel simulations), the signals at the output and – if available – at the input of the transmission channel under consideration can be used to predict quality. When the service is fully operational, measurements of the signal(s) and/or parameters can be used to predict quality during operation, in terms of quality monitoring. In this phase, it is distinguished whether monitoring is based on the degraded output signal which is available at normal service usage (so-called non-intrusive monitoring), or whether dedicated clean speech signals are transmitted over the channel and then quality is estimated on the basis of both, clean input and degraded output signal (so-called intrusive monitoring).

Algorithmic models are available for the planning and simulation phase, which predict an average overall quality rating – as it could be obtained in a subjective test with human participants. Such models might also be used for quality monitoring. However, they are limited as the reasons underlying a sub-optimal (predicted) quality remain obscure. Thus, in addition to predicting overall quality, service operators need estimations of the underlying technical reasons of sub-optimum quality. Identifying possible reasons provides diagnostic information for fixing problems, whereas the predicted overall quality helps to weight the severity of a problem.

Whereas diagnostic monitoring tools for narrowband speech communication services have been available for a long time [2][3], comparable tools are missing for super-wideband and fullband services. So far, only one parametric planning tool (fullband E-model [4]) and one intrusive signal-based prediction model for overall quality (POLQA [5]) have been developed, but experience in using them for monitoring quality is still scarce. No non-intrusive monitoring model is yet available, and algorithms for identifying technical causes in super-wideband or fullband have not yet been proposed.

It is the aim of this paper to contribute to filling this gap. For this purpose, we processed a speech database with coding and packet-loss degradations which are expected to be the major degradations encountered in real-life super-wideband and fullband services. Next, we tried to identify the type of codec used, as well as packet losses, on the basis of the degraded output signal only. We used the ground truth of the technical causes of impairments, as well as the estimated technical causes, to feed the parametric fullband E-model for estimating overall quality. The estimations are compared to intrusive signal-based overall quality estimations from POLQA, and an extension to the E-model is proposed for better capturing the impact of bursty packet loss on perceived quality. Section 2 reviews the current state-of-the-art of the POLQA and the fullband E-model. Section 3 presents the speech database used in the analysis. Section 4 describes the algorithms used for technical cause estimation and the results obtained, and Section

http://dx.doi.org/10.21437/Interspeech.2020-1125

5 the steps followed for estimating overall quality, including the extension to the fullband E-model. Section 6 discusses the results obtained, and points out limitations and future work.

2. Fullband speech quality prediction

Currently, there are two recommended standards for fullband speech quality prediction. The POLQA model recommended by the International Telecommunication Union [5] predicts overall listening quality on the basis of a degraded speech signal at the output of the communication channel and the corresponding clean, reference signal at the input of the same channel. After a pre-processing and alignment of both signals, POLQA predicts an average overall quality rating, a so-called Mean Opinion Score (MOS), from a perceptually-weighted difference between the degraded and a modified version of the clean speech signal. In contrast to earlier models such as PESQ [6], POLQA also has a fullband mode which enables fullband speech quality prediction. Details can be found in [5].

The second model is the fullband E-model [4] which predicts the overall quality experienced by a communication partner during conversations over a telephone channel exhibiting the characteristics as defined by the parameters of the model. So, unlike POLQA, the E-model predicts the quality of a conversation and not the listening quality. The main output of the model is a transmission rating *R* which can be transformed to MOS using an S-shaped monotonous relationship. In the narrowband version of the E-model [1], the transmission rating *R* ranges from 0 (being worst) to 100 (being optimal quality); in the fullband version [4], the maximum rating increases to 148. No significant difference between the maximum rating for super-wideband and fullband channels was found [7], so the model is expected to handle both situations.

The E-model assumes that impairments are independent from each other and can be quantified in terms of impairment factors on the transmission rating scale *R*. This can be done by subtracting the impairment factors from a maximal transmission rating:

$$R = 148 - I_{d,FB} - I_{e,eff,FB} \quad (1)$$

The term $I_{d,FB}$ represents impairments due to delay of the connection, restricting interactivity. These effects will not be addressed in the present paper (thus, we assume $I_{d,FB} = 0$), but first results are presented in [8].

The equipment impairment factor $I_{e,eff,FB}$ represents impairments caused by the codec as well as by the effects of randomly distributed packet-loss, derived using the codec-specific values for the equipment impairment factor at zero packet loss $I_{e,FB}$, the packet-loss probability P_{pl}, and the packet-loss robustness factor B_{pl}, see also [4] and [1]:

$$I_{e,eff,FB} = I_{e,FB} + \left(132 - I_{e,FB}\right) \cdot \frac{P_{pl}}{P_{pl}+B_{pl}} \quad (2)$$

In the current version of the fullband E-model, the calculation of the effective equipment impairment does not consider the burstiness of the packet-loss. However, in the narrowband version [1], bursty packet loss is accounted for by a so-called Burst Ratio *BurstR* defined as follows:

$$BurstR = \frac{\text{Average length of observed bursts}}{\text{Average length of bursts with random loss}} \quad (3)$$

Thus, the packet-loss is random when $BurstR = 1$ and bursty for $BurstR > 1$. With this parameter, the effective equipment impairment factor (without the indicator *FB*, as it is the narrowband case) is calculated by changing Eq. (2) as follows:

$$I_{e,eff} = I_e + (95 - I_e) \cdot \frac{P_{pl}}{\frac{P_{pl}}{BurstR}+B_{pl}} \quad (4)$$

The constant value of 95 instead of 132 is used because in the narrowband version the transmission rating *R* only varies between 0 and 100, instead of 0 and 148 for fullband. Values for $I_{e,FB}$ and B_{pl} are listed in [9].

3. Speech databases

We created a set of 18 clean mono speech files with 16-bit linear PCM and 44.1 kHz sampling rate. We coded these speech files both with the EVS codec in its super-wideband mode at 13.2 kbit/s, as well as with linear PCM at 44.1 kHz (fullband). On the coded speech files, we generated combinations of 6 packet-loss rates with 7 burst-ratios (see Table 1), resulting in 756 files for PCM and EVS each. To generate the packet-loss pattern we used a two-state Markov chain as described in [10]. We selected only generated patterns that did not deviate from the targeted *Ppl* by more than 1 percentage point. The packet-loss-affected speech files were finally decoded, upsampled to 48 kHz and – together with the clean speech files – used for the analysis.

Table 1: *List of packet-loss percentages (P_{pl}) and burst-ratios (BurstR) used in the experiment.*

Packet-Loss (%)	2.5, 5.0, 7.5, 10.0, 20.0, 30.0
Burst-Ratio	1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0

4. Non-intrusive technical cause estimation

Two types of technical causes are targeted in our analysis: The identification of the codec used and the estimation of packet loss. Both technical causes should be estimated from the degraded speech signal alone, i.e. in a non-intrusive way.

For the identification of the codec, the difference in signal bandwidth is exploited. The power spectral density is computed for each decoded signal to then obtain a value for the average power in the frequency ranges 0.5-3 kHz and 15-19 kHz, respectively. The logarithmic ratio of these two average powers is, in combination with a threshold, utilized as signal feature to distinguish between PCM and EVS-coded signals. We tested this non-intrusive approach on all of our test database, and it yielded an accuracy of 100%, with no misclassification.

For the non-intrusive estimation of packet loss, we follow the packet loss detection model presented in [11], but now extended to fullband speech. The model uses MFCCs (Mel-Frequency Cepstral Coefficients) as inputs. Additionally, to consider temporal effects, the first derivative is also included. The features are fed to a random decision forest which classifies each frame as either erroneous or unimpaired, resulting in an estimated error pattern for the speech file.

For this work, the resulting error pattern is used as input for a packet-loss rate prediction model. Since the described detection model is not able to find erroneous packets in silent segments, only active speech frames can be considered. As a result, the *Ppl* regarding the entire file (active and non-active frames) may be predicted by using a linear function:

$$P = -0.07 + 5.24\frac{N_L}{N_A} \quad (5)$$

where N_L is the number of detected erroneous frames and N_A the number of non-silent, active speech frames.

Figures 1 and 2 show the estimated vs. the actual packet loss rate for different burst ratios and for the EVS and the PCM codec, respectively. Whereas the estimation works very well for

the super-wideband EVS codec, it fails for the fullband PCM codec. We think that the training of the binary decision tree which classifies the features did not use enough PCM training data (only 38 out of 7080 training files). We expect that such a non-intrusive packet loss estimator requires training material to be available for all possible codecs used in the transmission. It is further interesting to note that the spread of the estimated P_{pl} as a result of burstiness increases with increasing P_{pl}.

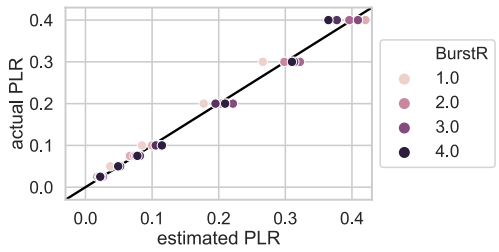

Figure 1: *Estimated vs. actual packet loss rate P_{pl} for different burst ratios BurstR for the EVS codec.*

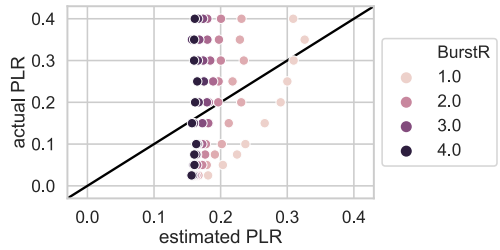

Figure 2: *Estimated vs. actual packet loss rate P_{pl} for different burst ratios BurstR for the PCM codec.*

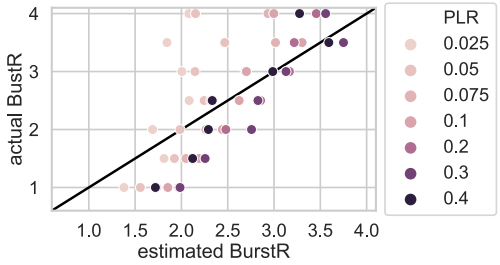

Figure 3: *Estimated vs. actual burst ratios BurstR for different packet loss rates P_{pl} for the EVS codec.*

Whereas the estimation of P_{pl} works quite fine for the EVS codec, it still results in larger errors for the estimated burst ration *BurstR*, as Figure 3 shows. Apparently, missed packet losses result in an over-estimation for small values of P_{pl} and an under-estimation for large P_{pl} values. The picture is slightly better, but similar for the PCM codec (omitted to save space).

5. Speech quality estimations

In the following, we will compare estimations of the fullband E-model with POLQA estimations regarding overall quality (MOS). For the fullband E-model, we will use the actual parameter values of the simulations (ground truth), as well as the parameter values obtained from the algorithms outlined in the previous section (estimated parameters). For the POLQA estimations, we used the SQuadAnalyzer software [12] in its fullband mode, and the POLQA version 3. Because the range of predicted MOS differs between POLQA and the E-model, the MOS prediction of POLQA was linearly scaled to the E-model MOS range of 1 to 4.5, and transformed to $I_{e,eff,FB}$ using Eq. 1. For the fullband E-model, we used $I_{e,FB} = 0$ for the PCM codec, and $I_{e,FB} =$ 17.1 for the EVS codec.

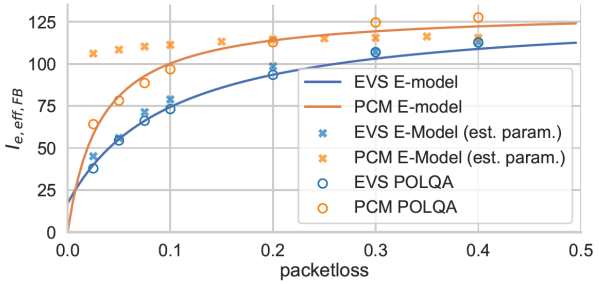

Figure 4: *$I_{e,eff,FB}$ values as predicted by POLQA and the fullband E-model for EVS (blue) and PCM (orange). Solid lines: E-model with actual parameter settings; crosses E-model with the estimated parameter settings. All predictions for BurstR = 1.0.*

Figure 4 shows the predictions of the fullband E-model in terms of $I_{e,eff,FB}$, with the actual settings for the codec and the packet-loss rate, when only considering random packet-loss. As expected, PCM shows a much lower robustness against packet-loss and even for very low packet-loss rates, linear PCM results in a higher effective equipment impairment factor $I_{e,FB,eff}$. When using the estimated parameter settings (orange crosses), the large estimation error for the PCM codec also results in large over-estimations of the $I_{e,eff,FB}$ value; the errors in the P_{pl} estimation for EVS (blue crosses) are negligible.

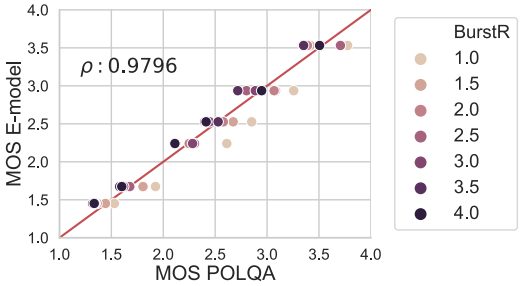

Figure 5: *Predicted MOS of POLQA versus predicted MOS of the fullband E-Model for EVS.*

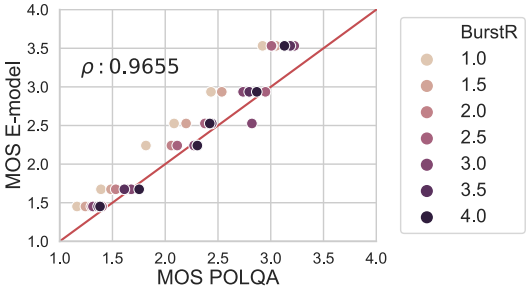

Figure 6: *Predicted MOS of POLQA versus predicted MOS of the fullband E-Model for linear PCM.*

Figure 5 shows the MOS of the EVS codec as predicted by POLQA versus the MOS predicted by the fullband E-model for different burst ratios *BurstR*, and the corresponding Pearson correlation ρ. The *BurstR*-levels are denoted by different colors. Because the fullband E-model in its current form does not

include the burst-ratio *BurstR* during the $I_{e,eff,FB}$ calculation, the predictions for a specific packet-loss rate produce the same MOS. The six distinct levels of packet-loss can be seen by the clusters in the E-model prediction. As can be seen by the colors of the scatter points, an increase in the burst-ratio results in a decreased MOS prediction from POLQA.

In contrast to that, the predictions for PCM (Fig. 6) have more variance between the burst ratios and the POLQA MOS *increases* with higher burst-ratios. Especially because of the high variance between the burst-rates, the RMSE for the E-model prediction (on the *R* scale [0;148]) is high with 12.36.

To accommodate for the burstiness of packet loss with the EVS codec, we modify the $I_{e,eff,FB}$ calculation similar to Eq. 4, however introducing a "burstiness robustness factor" B_{rf}:

$$I_{e,eff,FB} = I_{e,FB} + \left(132 - I_{e,FB}\right)\frac{P_{pl} - \frac{(1 - BurstR)}{B_{rf}}}{P_{pl} + B_{pl}} \quad (6)$$

Comparing Eq. 6 to Eq. 2, for random packet-loss ($BurstR = 1.0$) the extension $\frac{1-BurstR}{B_{rf}} = 0$, that means in this case the formula is the same as in the current fullband E-model. With increasing burst-ratio, the $I_{e,eff,FB}$ increases, and the penalization of the *BurstR* is independent of the packet-loss. The amount of penalization can be regulated with the B_{rf} value: Higher B_{rf} values signify higher robustness against burstiness, resulting in a smaller penalization of the burstiness. Also, with negative values for the B_{rf}, Eq. 6 is able to model codecs that increase in quality with higher burstiness, as it was observed for PCM in Fig. 6. The B_{rf} for each codec is fitted together with the B_{pl} value, and we obtained B_{rf} = 2.03 for the EVS and B_{rf} = -4.35 for the PCM codec.

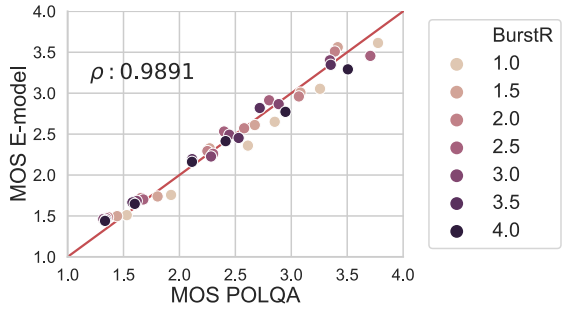

Figure 7: *Predicted MOS of POLQA versus the predicted MOS of modified fullband E-Model for EVS with the modified $I_{e,eff,FB}$ calculation from Eq. 6, using the actual simulation settings.*

As can be seen in Fig. 7 compared to Fig. 6, the consideration of burstiness using Eq. 6 shows a significant improvement of the E-model performance. The estimations of the E-model using the actual parameter values result in a very good agreement with the POLQA predictions. Using the estimated parameter values, Fig. 8 shows that the results are still in a very good agreement to the POLQA estimations. The RMSE of the extended E-model predictions with parameter estimations is quite low with 0.1178 on the 5-point MOS scale. This shows that using the 2-step diagnostic approach – i.e. first estimating codec and packet loss in a non-intrusive way, and then predicting overall MOS using the extended E-model, results in quite reasonable predictions compared to POLQA, even when the parameter estimations are not perfect. This finding is however limited to the EVS codec at one particular bitrate; whereas similar results have been obtained for the fullband PCM codec, other codec and bitrate settings still need to be investigated.

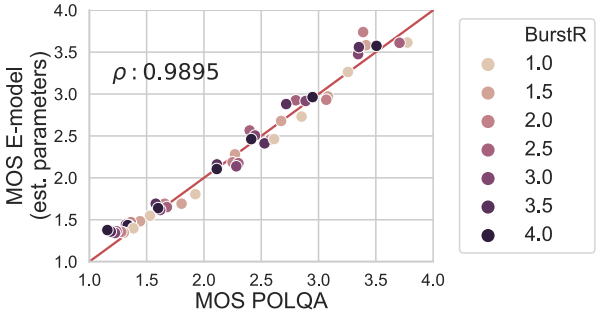

Figure 8: *Predicted MOS of POLQA versus the predicted MOS of modified fullband E-Model for EVS with the modified $I_{e,eff,FB}$ calculation from Eq. 6, using the estimated parameter settings.*

6. Discussion and future work

In this paper, we presented a first approach to non-intrusively monitor the quality of super-wideband and fullband speech transmission. Whereas standard monitoring requires both input and output signals to be available, our approach works in a purely non-intrusive way. By estimating technical causes as well as overall quality, it also serves the diagnosis of technical causes of impairments.

With the limited database consisting of two codecs only, we achieved a perfect identification of the codec used. This result should however be interpreted with care, as more codecs working at different bitrates may decrease the accuracy. The results obtained in [13] for the AMR-WB codec, however, make us confident that this could work with sufficient accuracy. The estimations for packet loss worked very well for the EVS codec, whereas they failed for the PCM codec. We consider the lack of PCM-coded training material for the binary decision tree as the major reason for the observed failure.

With the help of accurately estimated packet losses it becomes possible to also accurately estimate overall quality, using the fullband version of the E-model. Still, this model does not yet consider bursty packet loss. An extension to this type of loss improves the prediction performance of the non-intrusive approach significantly, and should also be considered in an update of the model standard.

Our approach was only tested on a very limited database, using two codecs and a limited amount of speech data consisting of sentences spoken in isolation (as they are required for using POLQA). If a non-intrusive approach is to be used in practice, it should also be able to cope with a variety of other codecs, and with speech material containing disfluencies, short backchannels, and so on. In addition, noise may be present which could impact codec identification and packet loss detection, as well as overall quality. Thus, more and more realistic speech data is necessary to test the robustness of the proposed approach. In addition, speech transmission channels are used in communication situations; thus, it would be interesting to analyze the effects of delay, and whether they are correctly covered by the fullband E-model.

Acknowledgements

This work was financially supported by the German Research Foundation DFG (grant number MO 1038/23-1).

7. References

[1] ITU-T Rec. G.107, "The E-model: a computational model for use in transmission planning," Geneva: Int. Telecomm. Union, 2015.

[2] ITU-T Rec. P:563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Geneva: Int. Telecomm. Union, 2004.

[3] D.-S. Kim, A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Techn. J.* 12(1), 2007, pp. 221-236.

[4] ITU-T Rec. G.107.2, "Fullband E-model," Geneva: Int. Telecomm. Union, 2019.

[5] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," Geneva: Int. Telecomm. Union, 2018.

[6] 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," Geneva: Int. Telecomm. Union, 2001.

[7] ITU-T Contr. SG12-C260, "Fullband extension of R-value," Source: Nippon Telegraph and Telephone Corporation (NTT), Geneva: Int. Telecomm. Union, 2018.

[8] T. Michael, G. Mittag, S. Möller, "Analyzing the Fullband E-model and Extending it for Predicting Bursty Packet Loss," in: *Proc. 12th Int. Conf. on Quality of Multimedia Experience (QoMEX 2020)*, 26-28 May 2020, IR-Athlone.

[9] ITU-T Rec. G.113 Amendment 2, "New Appendix V - Provisional planning values for the fullband equipment impairment factor and the fullband packet loss robustness factor", Geneva: Int. Telecomm. Union, 2019.

[10] A. Raake, "Speech quality of VoIP: Assessment and prediction," John Wiley & Sons, 2007.

[11] G. Mittag, S. Möller, "Single-ended Packet Loss Rate Estimation of Transmitted Speech Signals," in Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP), Anaheim, CA, USA, 2018, pp. 226-230.

[12] www.polqa.info.

[13] T. Hübschen, G. Mittag, S. Möller, G. Schmidt, "Signal-based Root Cause Analysis of Quality Impairments in Speech Communication Networks," in: *ITG Conference on Speech Communication*, DE-Oldenburg, 2018, pp. 4 pages.