INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Extending the Fullband E-Model Towards Background Noise, Bursty Packet Loss, and Conversational Degradations

Thilo Michael [1], *Gabriel Mittag* [1], *Andreas Bütow* [1], *Sebastian Möller* [1,2]

[1] Quality and Usability Lab, Technische Universität Berlin, Germany
[2] Speech and Language Technology,
German Research Center for Artificial Intelligence (DFKI), Germany
{thilo.michael|gabriel.mittag|sebastian.moeller}@tu-berlin.de

Abstract

Quality engineering of speech communication services in the full speech transmission band (0-20,000 Hz) is facilitated by the fullband E-model, a planning tool that predicts overall quality on the basis of parameters describing the setting of the service. We presented a first version of this model at Interspeech 2019, which has since then been standardized by the International Telecommunication Union in ITU-T Rec. G.107.2. Whereas that model was limited to predict the effects of speech codecs, random packet loss, and transmission delay, more realistic settings such as ambient background noise, bursty packet loss, as well as interactive conversational degradations could not be predicted. Based on the results of two new listening-only and conversational tests, we present an approach to extend the E-model to better predict these effects in the present paper. The results show that background noise effects at both sending and receiving side can be predicted well, whereas bursty packet loss predictions still have some limitations which result from the available database. Finally, approaches from conversational analysis help to better predict the effects of delay on conversational quality.

Index Terms: speech quality, quality prediction, transmission planning, E-model, background noise, packet loss, delay

1. Introduction

Not only due to the Covid-19 pandemic, IP-based speech and audiovisual communication services have become more important. As a result of the packet-based transmission technique, the traditional limitation to a narrow (300-3,400 Hz) audio transmission band has been overcome, and wideband (50-7,000 Hz), super-wideband (20-14,000 Hz) as well as fullband (0-20,000 Hz) transmission are commonly offered. Whereas this increase in transmitted audio band has improved overall quality by roughly 30 to 50% [1], the major impairments limiting perceived quality now stem from the used audio codec, packet loss, delay, as well as ambient noise at the sending and receiving side.

In order to plan speech communication services to reach an optimum quality for their users while avoiding over-engineering of required resources, parametric planning tools have been used since the old days of analog narrowband telephony. Such tools predict the overall conversational quality experienced by the conversation partners on the basis of assumptions about service components, as signals are not readily available during service planning. These components are described in terms of qualitative or quantitative parameters, such as the type of codec used, the level of ambient background noise, the attenuation and delay of a talker echo, etc. The parameters are then algorithmically merged to form a prediction of the overall conversational quality. The most widely used model is the E-model recommended by the International Telecommunication Union, ITU-T, in ITU-T Rec. G.107 [2]. This model is however limited to narrowband or wideband [3] speech transmission.

At Interspeech 2019, we presented a new version of this model which addresses both super-wideband and fullband transmission scenarios [1]. That model predicts the effects of fullband codecs (mostly EVS), random packet loss, as well as overall delay on the quality of a standard, not very interactive conversation. With a slight modification of the delay prediction, that model version has been standardized in ITU-T Rec. G.107.2 [4] since then. Still, the settings to be encountered in realistic services are not appropriately covered by that version: Speech communication occurs in acoustic environments where background noise – especially when transmitted with the full audio bandwidth – is a problem. Packet losses result from congestions in the network and occur in bursts rather than in an isolated random manner. And the effects of delay on conversational quality may differ between a standard private or an interactive business conversation [16]. As a result, the current version of the fullband E-model needs to be extended in order to be helpful for planning realistic conversational services.

This paper presents three approaches to overcome the mentioned limitations. First, we extend the predictions of the model by calculating a maximum signal-to-noise ratio limited by the background noise at both sending and receiving side. Second, we introduce a burstiness-specific degradation factor which allows to better predict the effects of non-random loss. Thirdly, we extend the predictions of the effects of delay by a metric which is derived from conversational analysis. The developed extensions are compared to the results of one listening-only and one conversational test. Section 2 presents the current version of the fullband E-model, and Section 3 presents the proposed amendments. Section 4 describes the details of the subjective tests used for the comparison, and Section 5 analyses the obtained results in comparison to the predictions. Conclusions and proposals for future work are given in Section 6.

2. Fullband E-model

The E-model predicts overall conversational quality based on the assumption that degradations which are in principle independent of each other can be expressed as so-called "impairment factors" on a "transmission rating scale". The transmission rating *R*, which serves as an index of overall conversational quality, can thus be calculated by subtracting individual impairment factors from a maximum transmission rating *Ro* which is given by the effective signal-to-noise ratio of the connection:

$$R = Ro - Id - Ie,eff \quad (1)$$

http://dx.doi.org/10.21437/Interspeech.2021-314

In this basic equation, *Id* is an impairment factor for degradations which occur delayed with respect to the transmitted speech signal (such as echo, or the degradation of the conversation due to pure delay), and *Ie,eff* is an impairment factor which describes the impairments caused by codecs under the effects of packet loss or discard. *R* can be transformed to an expected rating of users on a 5-point overall conversational quality scale as defined in ITU-T Rec. P.800 [5], MOS_{CQE}, using an S-shaped function defined in ITU-T Rec. G.107.2 [4]:

For $Rx < 0$: $MOS_{CQE} = 1$

For $0 < Rx < 100$: $MOS_{CQE} = 1 + 0.035\,Rx + Rx\,(Rx - 60) \cdot (100 - Rx) \cdot 7 \cdot 10^{-6}$ (2)

For $Rx > 100$: $MOS_{CQE} = 4.5$

with $Rx = {}^{R}/_{1.48}$. Eq. (2) considers that in a typical subjective experiment, the maximum average rating commonly being observed is around 4.5 on the MOS_{CQE} scale ranging from 1 to 5.

For the current fullband E-model [4], *Ro* is set to 148 and represents the quality advantage of fullband compared to narrowband speech communication services [2]. *Id* is calculated from the absolute delay between sender and receiver, *Ta*:
for $Ta \le 100$ ms:

$$Id = 0 \tag{3a}$$

for $Ta > 100$ ms:

$$Id = 1.48 \cdot 25\left\{(1 + X^6)^{\frac{1}{6}} - 3\left(1 + \left[\frac{X}{3}\right]^6\right)^{\frac{1}{6}} + 2\right\} \tag{3b}$$

and:

$$X = \frac{\log\frac{Ta}{100}}{\log(2)} \tag{4}$$

The impairment factor *Ie,eff* is derived using the codec-specific value for the equipment impairment factor at zero packet loss *Ie,* the packet loss probability *Ppl*, and the packet loss robustness factor *Bpl*:

$$Ie,eff = Ie + (132 - Ie) \cdot \frac{Ppl}{Ppl + Bpl} \tag{5}$$

Values for *Ie* and *Bpl* are listed in Appendix IV of ITU-T Rec. G.113 [6].

3. Proposed extensions

In the following paragraphs, we describe the extensions proposed to the standardized fullband E-model to better address the mentioned degradations.

3.1. Extension for background noise

Whereas the current fullband E-model sets *Ro* to a fixed value of 148, the corresponding narrowband E-model [2] calculates *Ro* on the basis of all circuit and background noise sources, assuming power addition of all equivalent electrical noise sources at the virtual 0 dBr point of the connection. We follow the approach of the narrowband E-model for first calculating the equivalent electrical power of all noise sources:

$$No = 10 \cdot \log_{10}\left(10^{Nc/10} + 10^{Nos/10} + 10^{Nor/10} + 10^{Nfo/10}\right) \tag{6}$$

In this equation, *Nc* represents the psophometrically weighted power of the circuit noise, and *Nfo* the electrical power of noise on the receiving line referred to the 0 dBr point; ITU-T Rec. G.107.2 assumes fixed values of -96 dBm0p for both sources. The equivalent electrical noise powers resulting from the acoustic ambient noise at the sending side with A-weighted power *Ps*, and at the receiving side with A-weighted power *Pr*, can be calculated using the send loudness rating *SLR* and the receive loudness rating *RLR*. For *Nos*, we follow Eq. (7) which is taken from the narrowband E-model:

$$Nos = Ps - SLR - Ds - 100 + 0.004 \cdot (Ps - OLR - Ds - 14)^2 \tag{7}$$

For *Nor*, we slightly modify the equation of the narrowband E-model to accommodate for the larger range of *R* values (0…148 in fullband compared to 0…100 for narrowband):

$$Nor = RLR - 147 + 1.12 \cdot Pre + 0.009 \cdot (Pre - 25)^2 \tag{8}$$

with the overall loudness rating *OLR* = *SLR* + *RLR*, *Ds* and *Dr* the differences in sensitivities of the used sending and receiving terminal devices for speech compared to background noises, and

$$Pre = Pr + 10 \cdot \log_{10}\left(1 + 10^{\frac{10 - LSTR}{10}}\right) \tag{9}$$

the effective room noise enhanced by the listener sidetone path with the corresponding loudness rating *LSTR*. From the value of *No* resulting from Eq. (6), the maximum value *Ro* is calculated by

$$Ro = 20 - 1.5 \cdot (SLR + No) \tag{10}$$

again slightly modifying the constant value to better cater for the extended range of *R* values in fullband.

3.2. Extension for bursty packet loss

In order to address the effects of bursty packet loss, we introduce the burst ratio *BurstR* into Eq. (5) which was also used in the narrowband E-model for describing the burstiness of the losses, see also [7]:

$$I_{e,eff,FB} = I_{e,FB} + \left(132 - I_{e,FB}\right)\frac{P_{pl} - \frac{(1 - BurstR)}{Brf}}{P_{pl} + B_{pl}} \tag{11}$$

This modification ensures that for random packet-loss ($BurstR = 1.0$) the extension $\frac{1 - BurstR}{Brf} = 0$, that means in this case the formula is the same as in the current fullband E-model. With increasing burst ratio, *Ie,eff,FB* increases independently of the packet loss rate *Ppl*. The amount of increase can be regulated with the factor *Brf* representing the robustness of the codec against bursty loss: higher *Brf* values, i.e. higher robustness against burstiness, result in a smaller penalization of the burstiness. Also, with negative values for the *Brf*, Eq. (11) is able to model codecs that increase in quality with higher burstiness. *Brf* values need to be derived together with the corresponding *Bpl* values for every codec, packet-size and packet-loss concealment used.

3.3. Extension for delay

It has been shown that the perceptual degradation caused by delay considerably depends on the interactivity of the conversation. As a consequence, parameters derived from conversational analysis have been proposed by Raake et al. [8]. This approach was taken up in the narrowband version of the E-model as an option to alternatively calculate *Id,* see also [15]:
for $Ta \le mT$:

$$Id = 0 \tag{12a}$$

for $Ta > mT$:

$$Id = 1.48 \cdot 25\left\{(1 + X^{6\cdot sT})^{\frac{1}{6\cdot sT}} - 3\left(1 + \left[\frac{X}{3}\right]^{6\cdot sT}\right)^{\frac{1}{6\cdot sT}} + 2\right\} \quad \text{(12b)}$$

and:

$$X = \frac{\log\frac{Ta}{mT}}{\log(2)} \quad \text{(12b)}$$

The parameter sT denotes the delay sensitivity of the participants, and mT reflects the minimal perceivable delay (in ms). Both parameters are dependent on the type of conversation that the prediction should reflect. In the narrowband E-model, sT is set to 1 and mT is set to 100 ms for a standard conversation, while in the case of lower delay sensitivity sT = 0.55 and mT = 120 ms, and in case of very low delay sensitivity sT = 0.4 and mT = 150 ms [2].

4. Subjective experiments

In order to quantify the impact of background noise at send and receive side on overall quality, a listening experiment has been carried out in a controlled laboratory environment in the frame of a Master thesis [9]. The room design followed ETSI EG 202 396-1 V1.2.2 [10] with respect to its size, equipment and noise floor (below 32 dB(A)). Four loudspeakers positioned at 1.25 m height in a square around the test participant were used to generate a rather diffuse background noise around the test participant.

In addition to the noise-free condition, three types of noise adjusted to levels of 55, 65 and 75 dB(A) have been considered, including speech-like babble which was however non-intelligible for test participants, as well as white and pink noise. Speech stimuli were presented via an open headset at approx. 74 dB(A) sound pressure level (SPL) and consisted of 12 double or triple sentences of approx. 10 s length, each recorded in fullband audio in a sound-insulated cabin from 2 female and 2 male speakers. The clean source files were degraded in the noisy environment by first calibrating the noise level at the position of the listener, then re-recording the noise with the headset by positioning it on a head-and-torso simulator HEAD Acoustics GmbH HMS II.3, and finally mixing the speech signal with the re-recorded noise at an active speech level of -26 dB relative to the overload point of the digital system, corresponding to an acoustic SPL of approx. 82 dB at the headset microphone. This procedure was repeated for all speech files using all available sending noise conditions (3 noise types x 3 noise levels), and a clean fullband, a wideband (following [12]) and a narrowband version (following [11]) of each file were added, resulting in 12 sending noise conditions (c1 to c12) with four stimuli each. The 48 stimuli were then presented either in quiet (background noise condition r1) or with pink noise of level 55, 65 or 75 dB(A) at the receiving side (background noise condition r2, r3 and r4).

25 participants reporting non-impaired listening rated the overall speech quality following the guidelines given in ITU-T Rec. P.800 [5] on a 5-point ACR scale. Before the start of the experiment, participants were informed about the purpose of the experiment, and listened to a range of 6 noise conditions to get accustomed to the quality levels to be expected in the experiment. They then carried out four separate sessions corresponding to the background noise conditions r1…r4, with randomized order of sending (c1…c12) and receiving (r1…r4) noise conditions. Participants were renumerated for their effort.

For addressing the effects of bursty packet loss and delay, a conversation test was carried out. This test followed either the Short Conversation Tests (SCT) or the Random Number Verification test (RNV) scenario, both defined in [13], to generate conversations of either standard (SCT) or high (RNV) interactivity. The test was conducted with a simulated conversational system which allowed the manipulation of the packet loss rate *Ppl* in three steps: 0%, 15% and 30%. The actual degradations included either bursty packet loss, pure delay, or combinations of packet loss and delay. The audio was coded with linear PCM and the lost packets were modeled with zero insertion and a burst-ratio of 4.0. Stereo headsets were used as listening and talking devices, as to minimize a potential acoustical echo. 27 participants (15 females, 12 males, aged 19-34 years) took part in the experiment and rated the overall quality on a 7-point extended continuous scale (ECS). To make the quality ratings comparable to the MOS range predicted by the fullband E-model, we converted the 7-point ECS ratings to the 5-point absolute category rating (ACR) as described by Köster et al. [14] and scaling the ratings to a range from 1 to 4.5.

5. Analysis of results

5.1. Background noise

The predictions for background noise using Eq.s (6) to (10) are compared to the results of the listening-only test in Fig. 1, for different levels and types of noise. The figure shows that the E-model predictions are well in line with the auditory test results when averaging over all noise types. The actual impact of babble noise in the auditory test seems to be less strong, and the impact of pink and white noise is stronger than predicted by the model. As the type of noise encountered in the conversation scenario is commonly unknown at the time of service-planning, this behavior seems acceptable for a planning tool.

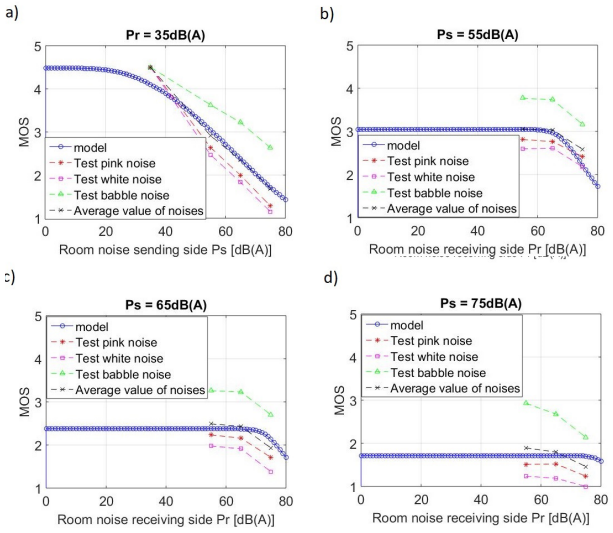

Figure 1: *Fullband E-model predictions using Eq.s (6) to (10). Panel a) shows the impact of sending side noise, whereas panels b) to d) address receiving side noise.*

5.2. Bursty packet loss

The predictions for the bursty packet loss using Eq. (11) are compared to the results of the conversation test in Fig. 2. The Bpl and Brf values for PCM with zero insertion packet-loss were derived using – in addition to the present conversation test – the results of a separate (unpublished) conversation test with 0%, 5%, 15%, 25%, and 35% PCM coded speech with zero

insertion packet loss with a burst ratio of 1.0. The resulting Bpl of 21.79 and the Brf of -6.9 lead to a more optimistic prediction than shown in [7]. While the current fullband E-model predicts pessimistic, high Ie,eff,FB values, our extension is able to model the burstiness of the packet loss and the burstiness robustness of the zero-insertion packet loss of the PCM coded speech.

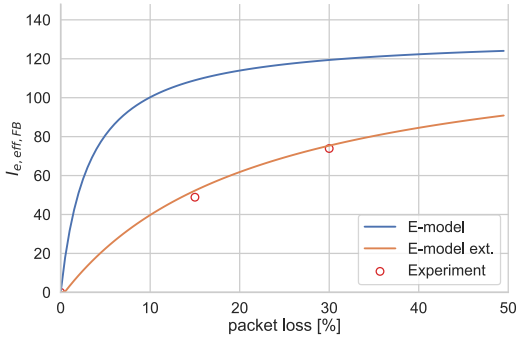

Figure 2: *Ie,eff,FB* values for different packet loss rates with *BurstR* of 4.0 as modeled by the current fullband E-model (blue), the extension using Eq. (11) (green) and the experimental data (red).

5.3. Delay

For the predictions with transmission delay, we split up the data from the conversation experiment into SCT and RNV scenario conversations and calculated the Id,FB values with the Eq.s (12a) and (12b). As recommended in [2], we set $sT = 0.4, mT = 150$ for the SCT predictions and $st = 0.55, mT = 120$ for the RNV predictions. Figures 3 and 4 show the predicted and actual MOS values for the combination of the three delay and the three packet loss levels for SCT and RNV conversations, respectively. The differences in interactivity between the SCT and RNV conversations can be modeled with the new equation for Id,FB.

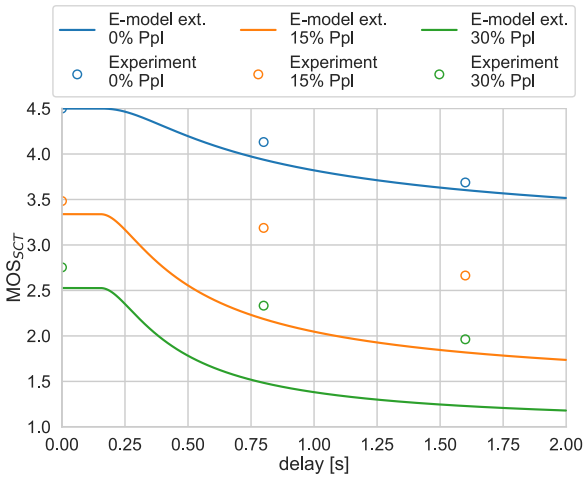

Figure 3: Fullband E-model predictions using the Eq.s (11) and (12) for various delay levels at 0%, 15%, and 30% packet loss. For experimental data as well as the delay parameters only SCT conversations were used.

While the predictions for conversations with only packet loss and conversations with only transmission delay are modeled by the E-model extensions, the combination of both degradations is predicted lower than the subjective MOS. For example, the quality ratings for conversations with 30% bursty packet loss are subjectively much less affected by transmission delay than it is predicted by the model, and similarly, the ratings for conversations with 1600ms transmission delay are subjectively less affected by the packet loss.

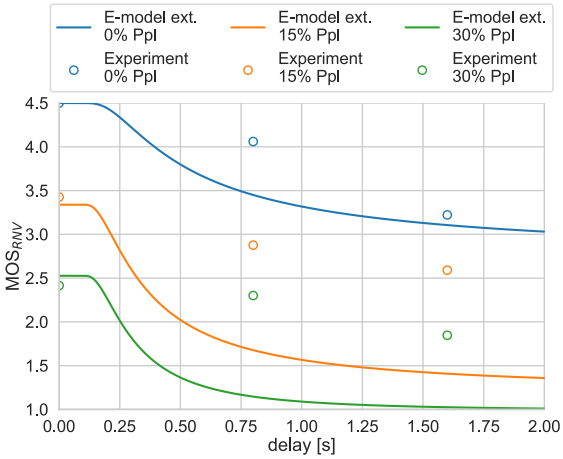

Figure 4: Fullband E-model predictions using the Eq.s (11) and (12) for various delay levels at 0%, 15%, and 30% packet loss. For experimental data as well as the delay parameters only RNV conversations were used.

6. Conclusions and future work

In this paper we presented three extensions for the fullband E-model, namely for conversations with background noise, with bursty packet-loss, and with transmission delay. We evaluated the extension for background noise with a listening-only test and the extensions for bursty packet loss and transmission delay with one conversation test, covering both degradations.

For the fullband E-model extensions that cover background noise, an excellent fit for the average of the noise conditions at the receiving and sending side could be obtained. The packet-loss extension using a burstiness-related formular could improve the prediction compared to the current version of the fullband E-model by including the burstiness of the packet-loss as well as the robustness of the codec against it. The delay extension is able to model the differences in the perception of the degradation during the highly interactive RNV conversations and the less interactive SCT conversations. When combining the extensions for bursty packet-loss and delay, the predictions are rather pessimistic. It should be noted, however, that slightly pessimistic worst-case predictions are in the nature of parametric planning tools; in case that the quality is slightly better than expected, this is for the benefit of the user, whereas a too optimistic prediction would put the planning process at risk.

In future work, we will confirm the results of the noise extension with a conversation test. We would like to validate the packet-loss extension with codecs like EVS that are used widely, and herewith derive more stable values for *Bpl* and *Brf*. In addition, we will validate the delay extension by carrying out conversation tests with more conversation scenarios with varying interactivity levels, and including the effects of talker echo. Lastly, we plan on improving the prediction results for combinations of these degradations; additional subjective conversation test results are necessary for reaching this aim.

7. References

[1] S. Möller, G. Mittag, T. Michael, V. Barriac, H. Aoki, " Extending the E-Model Towards Super-wideband and Fullband Speech Communication Scenarios", in *Proc. INTERSPEECH 2019,* Graz, Austria, pp. 3436-3440.

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

[3] ITU-T Recommendation G.107.1, *Wideband E-model*, Geneva: Int. Telecomm. Union, 2019.

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

[5] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*, Geneva: Int. Telecomm. Union, 1996.

[6] ITU-T Recommendation G.113, *Transmission impairments due to speech processing*, Geneva: Int. Telecomm. Union, 2007.

[7] T. Michael, G. Mittag, S. Möller, "Analyzing the Fullband E-model and Extending it for Predicting Bursty Packet Loss", in *Proc. IEEE QoMEX 2020*, 6 pages.

[8] A. Raake, K. Schoenenberg, J. Skowronek, S. Egger, "Predicting speech quality based on interactivity and delay". in *Proc. INTERSPEECH 2013*, Lyon, France, pp. 1384-1388.

[9] A. Bütow *Vorhersage der Qualität verrauschter Sprache bei vollbandiger Telefonie*, Master thesis, Quality and Usability Lab, Technische Universität Berlin, 2020.

[10] ETSI Guide EG 202 396-1 V1.2.2, *Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database*. European Telecommunications Standards Institute, Sophia Antipolis, 2008.

[11] ITU-T Recommendation P.48, *Specification for an Intermediate Reference System*, Geneva: Int. Telecomm. Union, 1988.

[12] ITU-T Recommendation P.341, *Transmission Characteristics for Wideband Digital Loudspeaking and Handsfree Telephony Terminals*, Geneva: Int. Telecomm. Union, 2011.

[13] ITU-T Recommendation P.805, *Subjective Evaluation of Conversational Quality*, Geneva: Int. Telecomm. Union, 2007.

[14] F. Köster, D. Guse, M. Wältermann, S. Möller, "Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech," in *Fortschritte der Akustik-DAGA*, 2015.

[15] T. Michael and S. Möller, "Interactivity-based Quality Prediction of Conversations with Transmission Delay," in *Proceedings of the 22nd International Conference SPECOM*, 2020, pp. 336–345, doi: https://doi.org/10.1007/978-3-030-60276-5_33.

[16] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE Journal on selected Areas in Communications*, vol. 9, no. 4, Art. no. 4, 1991.