Estimation of the Quality Dimension "Directness/Frequency Content" for the Instrumental Assessment of Speech Quality

Kirstin Scholz[1], *Marcel Wältermann*[2], *Lu Huo*[1], *Alexander Raake*[3], *Sebastian Möller*[3], *Ulrich Heute*[1]

[1] Institute for Circuit and System Theory, Christian-Albrechts-University of Kiel, Kiel, Germany
[2] Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany
[3] Deutsche Telekom Laboratories, TU Berlin, Berlin, Germany

ks@tf.uni-kiel.de

Abstract

The presented work aims at an instrumental method for the assessment of speech-quality of modern telecommunication networks that features two functions: the *prediction* of a system's overall speech-quality as well as the *analysis* of the overall speech-quality. The introduced method is based on the estimation of quality dimensions, each describing a particular characteristic relevant for the quality of speech signals. For narrowband telephone-speech three relevant dimensions have been identified: "directness/frequency content" , "continuity", and "noisiness". The estimates of the single dimensions allow for an analysis of speech quality while together forming a model for the overall quality.

This paper presents a method for the estimation of the quality dimension "directness/frequency content". The estimator is based on two parameters considered suitable to measure "directness/frequency content" of a speech signal $y(k)$ that has been transmitted over a telecommunication system. The estimates $\widehat{DF}$ of this dimension show correlations of $\rho > 0.93$ with the results of corresponding auditory tests.

Index Terms: instrumental assessment of speech quality, quality dimensions, frequency content.

1. Introduction

During the transmission of speech signals over modern telecommunication systems various disturbances and deformations of these signals occur. In general, these influences have an impact on the speech quality and, thus, the system quality perceived by a system's user.

The *integral* perceived speech-quality can be described in terms of the mean-opinion score (MOS). Well-defined methods exist for the auditory assessment of the MOS value of a system. Mostly, an absolute-category rating in a listening-only situation is used. Only the quality statements obtained from human listeners within such auditory tests can be considered to be valid. But auditory experiments are expensive and time-consuming. Thus, there is a need for a more efficient assessment of speech quality. This is achieved by an instrumental *estimation* of speech quality. Such instrumental methods are considered reliable if their MOS estimates are highly correlated with the corresponding quality statements obtained in auditory experiments. Instrumental, signal-based methods for the assessment of a system's integral speech-quality are, e.g., the ITU-T standard P.862 [1] or the TOSQA model [2].

However, methods providing a single MOS estimate to describe the integral speech quality of a system do not allow to *analyze* quality. The same MOS might be provided for two signals which are perceived to sound differently by human listeners. Thus, a quality analysis might be useful for a better distinction between the quality of various signals. Furthermore, such an analysis might provide starting points for improving a system's quality. For the purpose of analyzing speech quality a more differentiated quality measure is needed.

As described in [3] our work aims at an instrumental method for the prediction and the analysis of speech-quality based on the assessment of *quality dimensions*. Every single quality dimension describes a different quality-relevant characteristic of a speech signal. An estimate for the MOS of a system will be obtained by a suitable combination of the judgements of the different dimensions. For narrowband telephone-speech three relevant dimensions have been identified [4]: "directness/frequency content", "continuity", and "noisiness". Section 2 gives an overview of these three quality dimensions.

This paper deals with the development of an estimator for the dimension "directness/frequency content". Thus, section 3 discusses the results of the auditory tests regarding this dimension as well as so-called dimension parameters considered suitable for the estimation of this dimension. Section 4 presents an estimator for "directness/frequency content" based on two dimension parameters and discusses the obtained dimension estimates $\widehat{DF}$. Section 5 gives a conclusion of this paper as well as an outlook to our future work.

2. Quality dimensions

In [4] three dimensions have been identified as relevant for the assessment of the quality of a speech signal $y(k)$ that has been transmitted over a modern telecommunication network with a bandlimition to a maximum frequency of 4kHz:

1. "Directness/frequency content": It is assumed that this dimension reflects the quality-relevant influence of a transmission system's frequency response on the signal $y(k)$.
2. "Continuity": This dimension is related to influences of the signal form of $y(k)$ in the time domain, e.g., due to interruptions caused by packet loss or due to musical tones caused by noise reduction.
3. "Noisiness": This dimension covers the perceptual influence of the signal $y(k)$ due to noiselike disturbances, e.g., background noise or circuit noise.

Expressing the integral quality (MOS) of a speech signal $y(k)$ in terms of a weighted linear combination of these three quality di-

mensions leads to a model for the overall quality that covers about 90% of the total variance of the speech quality judgements [4].

3. "Directness/frequency content"

3.1. Results of the auditory tests

To identify and to characterize perceptual quality dimensions of modern telecommunication networks, multidimensional analyses (MDA) were carried out in [4] using twenty-eight stimuli. The stimuli were obtained by processing two German sentences spoken by one male and one female speaker with fourteen test conditions each. The test conditions represent different telecommunication-system configurations and comprehend simple codecs, the use of hands-free terminals (HFT), disturbances due to background noise and circuit noise, the utilization of noise-reduction algorithms, the occurrence of packet loss as well as the application of bandpasses. All obtained stimuli are narrowband-speech signals.

The semantic differential experiment conducted in [4] evinces a high correlation of "directness/frequency content" with five attribute pairs: *distant-close* and *indirect-direct* (generic term: "directness") as well as *thin-thick*, *muffled-not muffled* and *dark-bright* (generic term: "frequency content").

Mainly two groups of stimuli obtain high negative ratings of "directness/frequency content":

1. Stimuli with a suppression of high and low frequency components due to a distinct bandlimitation of a system's frequency response.
2. HFT-stimuli that are processed by systems whose magnitude frequency responses are characterized by the occurrence of peaks and notches (comb filter effect).

A high positive rating of the dimension "directness/frequency content" is typical for stimuli processed by systems whose frequency responses feature neither a significant bandlimitation nor a conspicuous comb filter effect. Among this group are:

1. Stimuli that are generated by use of simple codecs.
2. Stimuli characterized by the use of a simple codec and the addition of circuit noise.

The results of the auditory tests indicate that the rating of the dimension "directness/frequency content" of a speech signal $y(k)$ is related to characteristics of the frequency response of the system by which $y(k)$ has been processed. This hypothesis is corroborated by both the attribute pairs characterizing this dimension and the features of the stimuli groups showing different ratings of "directness/frequency content".

3.2. Dimension parameters

According to section 3.1, for estimating "directness/frequency content" a model is needed describing characteristics of a system's frequency response $H'(e^{j\Omega})$. In this work $H'(e^{j\Omega})$ is given in terms of the gain function $G'(\Omega) = 20 \cdot lg|H'(e^{j\Omega})|$ dB. An example for $G'(\Omega)$ of a used test condition is given in Fig. 1. This test condition named H (the test condition's name "H" stands for "hands-free terminal") includes a G.711 codec, a hands-free terminal and the application of a typical handset filter on the sender side [4]. The following five parameters of $G'(\Omega)$ are used to model the characteristics of a system's frequency response:

- *Bandwidth and center of gravity of* $G'(\Omega)$: These parameters indicate the frequency components being transmitted by a system.

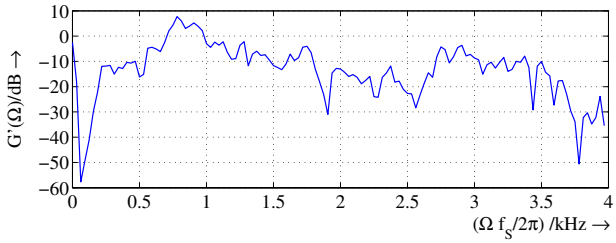

Figure 1: *Gain function* $G'(\Omega)$ *of the test condition H.*

- *Slope of* $G'(\Omega)$: The slope characterizes the ratio of the frequency components being transmitted by a system.
- *Depth and width of peaks and notches of* $G'(\Omega)$*:* These parameters are assumed to be related to the occurrence of reverberations.

All these parameters are considered to have an impact on the dimension "directness/frequency content" and, thus, are entitled as *dimension parameters* in the following.

It is the aim of this work to analyze the relations between the dimension parameters and the rating of "directness/frequency content" of systems to be able to establish a dimension estimator.

At present time, the judgements of "directness/frequency content" for twenty-eight speech signals are available for this purpose. Hence, to avoid an over-parameterization of the estimator, this paper concentrates on the analyses of the two parameters and their influence on "directness/frequency content" that appear to have the greatest impact on this dimension: the bandwidth and the center of gravity of $G'(\Omega)$.

4. Dimension estimator

This section deals firstly with the estimation of the dimension parameters bandwidth and center of gravity of $G'(\Omega)$. Afterwards the dimension estimator is presented.

Before the actual estimation of the dimension parameters, the determination and a preprocessing of $G'(\Omega)$ is carried out. During the preprocessing, $G'(\Omega)$ is modified such that the dimension parameters of the modified gain function are optimally suitable for building a linear estimator for "directness/frequency content".

4.1. Determination and preprocessing of $G'(\Omega)$

4.1.1. Determination of $G'(\Omega)$

The gain function $G'(\Omega)$ of a transmission system is given by:

$$G'(\Omega) = 20 \cdot \lg |H'(e^{j\Omega})| = 20 \cdot \lg \left[\frac{|\Phi_{xy}(e^{j\Omega})|}{|\Phi_{xx}(e^{j\Omega})|} \right]. \quad (1)$$

$\Omega = 2\pi \frac{f}{f_S}$ gives the frequency f normalized to the sampling frequency f_S. $\Phi_{xx}(e^{j\Omega})$ represents the power-density spectrum of the input signal to the system, $x(k)$, and $\Phi_{xy}(e^{j\Omega})$ the cross-power-density spectrum of $x(k)$ and $y(k)$.

4.1.2. Determination of $G(\theta)$

A signal analysis on the Bark scale has proven to be useful when modeling the auditory perception of signals (e.g., [1, 2]). Thus, according to the method for modeling the influence of bandpasses on speech quality presented in [5], the bandwidth of a gain function

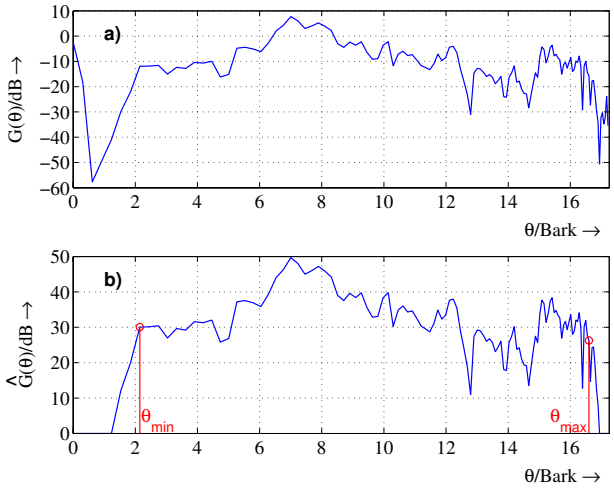

Figure 2: *a) Gain function $G(\theta)$ of the test condition H. b) Modified gain function $\widehat{G}(\theta)$ of the test condition H and the limits of the critical-band rate interval $\Delta\theta$, θ_{min} and θ_{max}.*

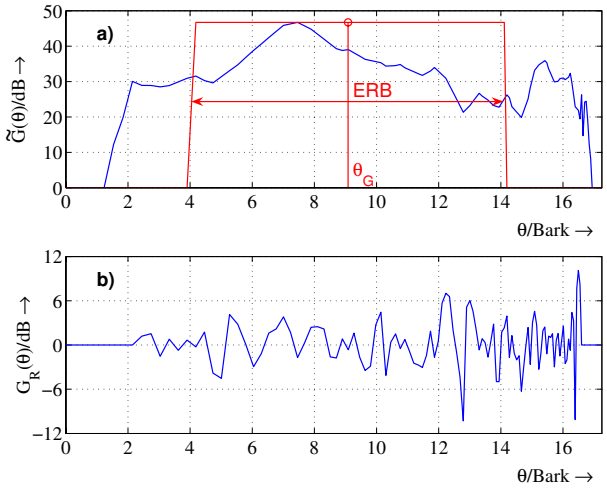

Figure 3: *a) Smoothed gain function $\widetilde{G}(\theta)$ of the test condition H with the values of ERB and θ_G. b) Estimated course of the peaks and notches $\widehat{G}_R(\theta)$ of the test condition H.*

is determined in terms of the critical-band rate θ. In this paper, the center of gravity is also given in terms of θ. For this purpose, $G'(\Omega)$ is transformed to $G(\theta)$ using the following relation between the frequency f and the critical-band rate θ [6]:

$$\frac{\theta}{\text{Bark}} \approx 13 \arctan\left(0.76\frac{f}{\text{kHz}}\right) + 3.5 \arctan\left[\left(\frac{f}{7.5\text{kHz}}\right)^2\right]. \quad (2)$$

The result of the transformation of the gain function $G'(\Omega)$ of the test condition H, shown in Fig. 1, to the critical-band rate scale is shown in Fig. 2a).

4.1.3. *Restriction to the critical-band rate interval $\Delta\theta$*

During the subsequent estimation of the dimension parameters only the critical-band rates within the interval $\Delta\theta = [\theta_{\min}, \theta_{\max}]$ are considered. The interval limits are determined in two steps:

1. Modification of $G(\theta)$ pursuant to:

$$\widehat{G}(\theta) = \begin{cases} \max\{G(\theta) + ST, 0\}, & \text{for } \theta \in [1.5, 17.0], \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

   The parameter ST is chosen such that based on the dimension parameters of $\widehat{G}(\theta)$ an optimal linear dimension estimator can be established. In this paper, ST=42 dB holds.

2. Choice of the interval limits $\theta_{\min}$ and $\theta_{\max}$ according to:

$$\widehat{G}(\theta < \theta_{\min}) < 0.5 \cdot \max\{\widehat{G}(\theta)\}, \quad (4)$$

$$\widehat{G}(\theta > \theta_{\max}) < 0.5 \cdot \max\{\widehat{G}(\theta)\}. \quad (5)$$

The modified gain function $\widehat{G}(\theta)$ of the test condition H as well as the corresponding limits of the critical-band rate interval $\Delta\theta$, $\theta_{\min}$ and $\theta_{\max}$, are depicted in Fig. 2b).

4.1.4. *Decomposition of $\widehat{G}(\theta)$*

Within the interval $\Delta\theta$, $\widehat{G}(\theta)$ is decomposed according to:

$$\widehat{G}(\theta) = \widetilde{G}(\theta) + \widehat{G}_R(\theta). \quad (6)$$

$\widetilde{G}(\theta)$ describes a smoothed curve of $\widehat{G}(\theta)$, and $\widehat{G}_R(\theta)$ the estimated course of the peaks and notches of $\widehat{G}(\theta)$. The result of the decomposition of $\widehat{G}(\theta)$ into $\widetilde{G}(\theta)$ and $\widehat{G}_R(\theta)$ for the test condition H is shown in Fig. 3.

4.2. Extraction of dimension parameters

The dimension parameters bandwidth and center of gravity of $G'(\Omega)$ are determined based on the corresponding smoothed gain function $\widetilde{G}(\theta)$.

4.2.1. *Bandwidth ERB*

According to [5], the bandwidth of $G'(\Omega)$ is given in terms of the equivalent rectangular bandwidth (ERB) of $\widetilde{G}(\theta)$:

$$ERB = \frac{\text{area}\{\widetilde{G}(\theta)\}}{\max\{\widetilde{G}(\theta)\}}. \quad (7)$$

4.2.2. *Center of gravity θ_G*

The center of gravity θ_G of a gain function $G'(\Omega)$ is determined as the center of gravity of $\widetilde{G}(\theta)$:

$$\theta_G = \frac{\int_{\theta_{\min}}^{\theta_{\max}} \widetilde{G}(\theta) \cdot \theta \; d\theta}{\int_{\theta_{\min}}^{\theta_{\max}} \widetilde{G}(\theta) \; d\theta}. \quad (8)$$

The values of the parameters, $ERB = 10.21$ Bark and $\theta_G = 9.08$ Bark, for the test condition H are depicted in Fig. 3a).

4.3. Dimension estimate $\widehat{DF}$

Fourteen stimuli from [4] are used for training and testing the estimator for "directness/frequency content", respectively. The sets of stimuli used for training and testing each contain seven stimuli uttered by the female speaker and seven stimuli uttered by the male speaker. Furthermore, both sets of stimuli cover all used test conditions.

The model for estimating the "directness/frequency content" is based on the parameters ERB and θ_G. Here, the limited number

of speech samples available for training the dimension estimator allows only for a linear model. Thus, by means of linear regression the following model for estimating the "directness/frequency content" has been revealed for the training stimuli:

$$\widehat{DF} = -20.5865 + 0.2466 \cdot \frac{ERB}{\text{Bark}} + 1.8730 \cdot \frac{\theta_G}{\text{Bark}}. \quad (9)$$

For the purpose of evaluating the reliability of the model given in Eq. (9), the predictor has been applied for the estimation of all *three* dimensions of both groups of stimuli, the training stimuli and the test stimuli. For both groups, Table 1 gives the correlation of $\widehat{DF}$ with the results of the auditory tests and the root mean-square error of the estimate $\widehat{DF}$ for each single quality dimension.

Table 1: *Correlation of $\widehat{DF}$ with the results of the auditory tests and the root mean-square error of the prediction for both groups of stimuli, training and test stimuli, and all three quality dimensions.*

	Training		**Test**	
dimension	*correlation*	*RMSE*	*correlation*	*RMSE*
direct./freq.cont.	0.9635	0.2700	0.9356	0.3331
continuity	-0.1032	1.3699	-0.1804	1.5276
noisiness	0.0610	1.3577	0.0731	1.2659

For both groups of stimuli, the estimates $\widehat{DF}$ show a high correlation with the results of the auditory tests for the dimension "directness/frequency content". In contrast, the correlations of $\widehat{DF}$ with the results of the auditory tests for the two other dimensions, "continuity" and "noisiness", are negligible. This result shows that the dimension estimator fulfills an important prerequisite for a dimension-based quality model: The linear combination of dimension parameters used to estimate $\widehat{DF}$ does not allow the other perceptual dimensions to be estimated.

Fig. 4 shows the estimates $\widehat{DF}$ plotted versus the results of the MDA for the group of test stimuli. This plot reveals the goodness of the presented dimension estimator. All stimuli are found grouped around the bisecting line. This means that corresponding to the auditory test the estimates $\widehat{DF}$ of the HFT-stimuli (H_f, HN_f, $HNR1_m$, $HNR2_m$) as well as the stimulus with a distinct bandlimitation (BP_m) are found at the negative end of the $\widehat{DF}$-axis. Stimuli

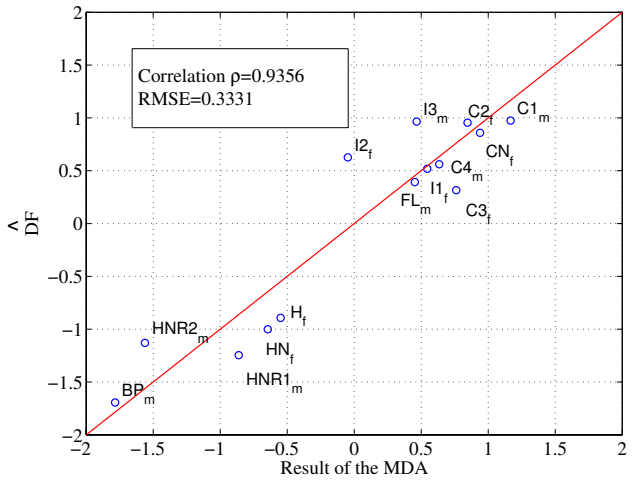

Figure 4: *The estimates $\widehat{DF}$ plotted versus the results of the MDA for the group of test stimuli.*

generated by use of simple codecs ($C1_m$, $C2_f$, $C3_f$, $C4_m$) and the addition of circuit noise (CN_f) obtain high estimates $\widehat{DF}$.

The variance of the overall quality judgements that is covered by the estimate $\widehat{DF}$ has a value of 0.5035 for the training stimuli and a value of 0.3630 for the test stimuli. These low values for the correlation of the estimates $\widehat{DF}$ with the results of the auditory test for the MOS agree with the finding of [4].

5. Conclusion and outlook

This paper has discussed the properties of the quality-relevant dimension "directness/frequency content". A method for the estimation of this dimension has been presented. The provided dimension estimates $\widehat{DF}$ show a correlation of $\rho > 0.93$ with the results of corresponding auditory tests. The estimate $\widehat{DF}$ is based on two parameters of a transmission system's gain function $G'(\Omega)$: the bandwidth ERB and the center of gravity θ_G of $G'(\Omega)$.

In our future work, more stimuli will be analyzed to further evaluate the influence of the parameters ERB and θ_G on "directness/frequency content". In addition, the influence on this dimension by the dimension parameters mentioned in section 3.2 that, so far, are not included in the dimension estimator, will be analyzed. Based on the results of these studies, the dimension estimator will be improved. For a systematic study of the influence of all five dimension parameters on "directness/frequency content", an idealized model of this dimension will be used that allows for the processing of speech signals using systems with defined values of the dimension parameters.

Estimators for the other quality dimensions "continuity" and "noisiness" are under development. In our future work, all three estimators will be combined to form a model for a system's MOS.

6. Acknowledgements

The authors would like to thank the "Deutsche Forschungsgemeinschaft" (DFG) for the financial support of the present work under the grants HE 4465 and MO 1038.

7. References

[1] ITU-T Rec. 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." ITU-T, CH-Geneva, 2001.

[2] Berger, J., "TOSQA - Telecommunication Objective Speech-Quality Assessment." ITU-T, Contr. COM-12-34, CH-Geneva, 1997.

[3] Heute, U.; Möller, S.; Raake, A.; Scholz, K.; Wältermann, M., "Integral and Diagnostic Speech-Quality Measurement: State of the Art, Problems, and New Approaches". In: Proc. Forum Acusticum 2005, H-Budapest 2005, pp. 1695-1700.

[4] Wältermann, M.; Scholz, K.; Raake, A.; Heute, U.; Möller, S., "Underlying Quality Dimensions of Modern Telephone Connections". Accepted for the International Conference on Spoken Language Processing 2006, Pittsburgh, PA, 2006.

[5] Raake, A., Speech Quality of VoIP - Assessment and Prediction. UK-Chichester, West Sussex: Wiley, 2006.

[6] Zwicker, E.; Fastl, H.: Psychoacoustics: Facts and Models. Springer, D-Berlin, 1999.