euronoise

Estimation Model for the Speech-Quality Dimension 'Noisiness'

Lu Huo[a], Marcel Wältermann[b], Ulrich Heute[a] and Sebastian Möller[b]

[a]Institute for Circuit and System Theory, Christian-Albrechts-University of Kiel, Kaiserstrasse 2, 24143 Kiel, Germany
[b]Deutsche Telekom Laboratories, Berlin Institute of Technology, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
lhu@tf.uni-kiel.de

A signal-based model is introduced which intends to predict integral speech quality along with diagnostic information in the context of noise. Since "noisiness" is one of the most important perceptual dimensions for the quality of transmitted speech, this measure constitutes one building block for an attribute-based speech quality measure which should be capable to cover larger-scale degradations introduced by speech transmission systems.
The "noisiness" estimation model is based on the prediction of so-called sub-dimensions, reflecting the relevant perceptual dimensions for noisy conditions. By means of auditory experiments and a subsequent multidimensional analysis, three sub-dimensions "speech contamination", "(perceived) additive-noise level", and "noise coloration" could be revealed. A two-step prediction model is applied for quality estimation: Firstly, the sub-dimensions are predicted by the combination of instrumental measures. Secondly, "noisiness" is estimated by combining the extracted sub-dimensions. A correlation of $\rho \geq 0.94$ can be achieved.

1 Introduction

State-of-the-art instrumental assessment methods for the speech-transmission quality (e.g., PESQ [1]) predict the so-called mean opinion score (MOS) quite accurately for traditional narrowband speech (300-3400 Hz). However, such methods do not provide diagnostic information of the speech quality. Such information, however, can give useful insight into the sources for a decrease in quality and, thus, are desirable for system or network developers and maintainers.

In a current research project, we aim at developing such a diagnostic model for speech-quality prediction. Integral quality is predicted on the basis of a signal-based measurement of perceptual attributes. In [2], it has been shown that the three, mutually orthogonal dimensions "directness/frequency content", "continuity", and "noisiness" are essential for traditional telephone-speech quality.

The project roadmap thus encompasses the development of three corresponding dimension estimators and the derivation of a suitable mapping for integral-quality prediction. In the present study, an estimator for the dimension "noisiness" is introduced. The basis for this measure is the prediction of perceptual and orthogonal dimensions *in the context of noise*. These dimensions can be understood as *sub-dimensions* (SDs) , in contrast to the *global* dimensions mentioned above that were derived in the context of diverse types of degradations.

A two-step estimation model is chosen here: the condition's position on SDs is firstly predicted by means of signal parameters correlating with these perceptual dimensions. Secondly, the estimated SD coordinates are mapped onto the sought integral quality scores.

The revelation of these SDs is done by means of auditory experiments and multidimensional analysis techniques. In order to study the relation between the SDs and integral quality, overall quality scores were collected in a further test. Both experiments are described in Section 2. In Section 3, signal-based *dimension parameters* and extraction methods are introduced which are intended to capture the effects the auditory test revealed. By combining these parameters to *dimension estimators*, the SD scores as well as the integral quality can reliably be predicted (Section 4). Finally, conclusions and discussions are provided in Section 5.

2 Auditory Tests

For the development process of the two-step prediction model, auditory test results are required that form the target values to be estimated. According to our approach, two auditory tests were conducted. In order to reveal the perceptual space, a similarity-scaling experiment with subsequent Multidimensional Scaling (MDS) was carried out. The basic idea of MDS is to translate the rated dissimilarity of each stimuli pair into a corresponding distance (the more dissimilar two stimuli are, the larger the distance). In this way, a point configuration can be determined representing the stimuli in an L-dimensional space. The dimensionality L is derived on the basis of both statistical fit parameters like the *Stress* and the ability for an interpretation [3].

Since the number of stimuli is large in this study, the so-called "Sorting Task" was employed as an efficient method of similarity scaling. Here, the participants were asked to group similarily sounding stimuli into common bins. As a similarity measure, the frequency of occurrence of two stimuli in a common group is counted over all participants, as described in [4].

The second test consisted of a scaling of integral speech quality on a continuous 5-point scale with overflow ranges (cf. [5]) principally following the guidelines of ITU-T Rec. P.800.

For the experiments, a set of speech samples was produced which covers a wide range of distortions in the "noisiness" domain according to the findings in [2]:

- Additive noise: (White) electric circuit noise (band-limited to 300-3400 Hz), (white) noise induced on subscriber lines (cf. [5]), and ambient noise, acoustically induced at send side ("Hoth"-shaped white noise, noise of a car's driver cabin at 100 km/h, pneumatic hammer, cafeteria noise).

- Multiplicative noise: Noise stemming from adaptive differential pulse-code modulation coding (ADPCM, see ITU-T Rec.G.726) and the modulated noise reference unit (MNRU, see ITU-T Rec. P.810).

In each case, different noise levels were applied. A majority of the speech samples were produced by means of a realistic telephone simulation tool [5]. A G.711-coded condition was included as a typical reference in standard telephony. In total, 69 conditions were considered for each of four speakers (2f, 2m) in the sorting task (one separate session for each speaker). Due to experimental effort, only 42 out of the 69 conditions were rated

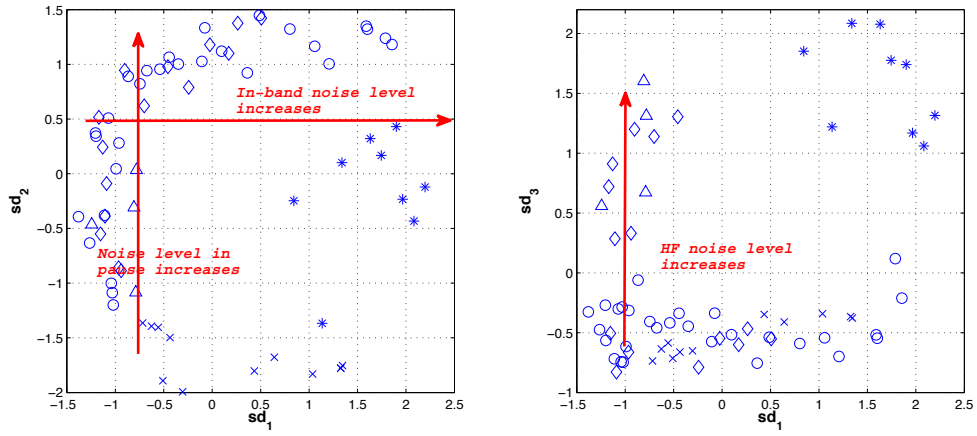

Figure 1: *3D configuration plots for one female speaker. Left: sd_1 vs. sd_2 Right: sd_1 vs. sd_3. Cross $\times$: signal-corr. noise; star $*$: real backg. noise (car, hammer, babble); diamond $\diamondsuit$: noise with both LF (below 3400 Hz) and HF (above 3400 Hz) components; triangle $\triangle$: HF noise only; circle $\circ$: the other cases.*

in the integral quality experiment; however, all *types* of degradation were included here. A number of 20 listeners (5 f, 15 m) took part in the experiment, aged between 18 and 47 years ($\varnothing \approx 26.7$ years). They listened to the stimuli through a standard handset, equalized according to the standard receive characteristic IRS-rec. The experiments took place in a room conforming to the requirements given in ITU-T Rec. P.800.

A dimensionality of $L = 3$ results in an acceptable data fit (Kruskal's $Stress_1 < 0.1$, cf. [3]) and provides a plausible picture of the data. Fig. 1 depicts two planes of the resulting point configuration for one female speaker. In the remainder, we focus on an axial interpretation of the structure:

- SD1: The first SD is labeled with *speech contamination*. In its positive direction, the energy of both the additive and multiplicative noise within the speech-relevant spectral band (limited to 300-3400 Hz) increases. Significant noise energy beyond this band (mainly stemming from the subscriber-line's noise-floor) is not captured here. Their respective points are located on the negative end of this axis.
- SD2: The second SD is labeled *(perceived) additive-noise level*, as it is highly correlated with the energy of additive noise.
- SD3: The third SD is labeled *noise coloration*. Along this dimension, the strongly colored noises such as realistic background (BG) noises and the very "bright" sounding subscriber-line's noise-floor can be distinguished from the rest.

Comparing the point configuration of the different speakers (each of which are derived from the data of a dedicated test session), it can be stated that the point coordinates are mostly stable, except for the realistic BG conditions. Obviously, the participants were uncertain about the similarity of the BG stimuli with regard to each other and to the other (white or shaped white noise) conditions. This may be due to their strongly differing composition and information content.

3 Dimension Parameters

Fig. 2 shows the flow diagram of signal processing used for feature extraction. Firstly, a preprocessing is deployed to suppress the influence from the distortions other than noise. Furthermore, it provides suitable the internal representation of signals for parameter extraction. Five parameters are extracted for SD estimation according to the following ideas:

- As SD1 is assumed to be determined by the degree of speech distortion by "in-band" noise (300-3400 Hz), measures should be developed to capture either the speech distortion or the noise level in the speech activity. If the additive noise is constantly present throughout the stimuli, the LF noise level n_{lf} is measured in speech pauses and below 3400 Hz. If no substantial noise is found in speech pauses, the weighted cepstral distance d_{cep} is used to measure the "fine" speech distortion such as signal-correlated noise [6].
- For SD2, the noise level is measured as n_p in speech pauses.
- Without BG noise, SD3 seems to be solely influenced by the very "bright" noise that can be distinguished from the rest. As a result, the gravity center of frequency $f_{c,n}$ of noise [7] is used to describe the noise-energy distribution and the HF (above 3400 Hz) noise level n_{hf} is also measured.

3.1 Preprocessing

Here, the preprocessing is described in detail.

3.1.1 Time-frequency Analysis and Gain Equalization

The spectra of both the clean and the degraded signal, notated as $X(\mu, i)$ and $Y(\mu, i)$, are calculated by a short-time Fourier transformation (STFT). Here, μ indicates the μ-th frequency bin and i indicates the i-th frame, where $\mu = 0, 1, \cdots, M-1$ and $i = 0, 1, \cdots, L-1$. The

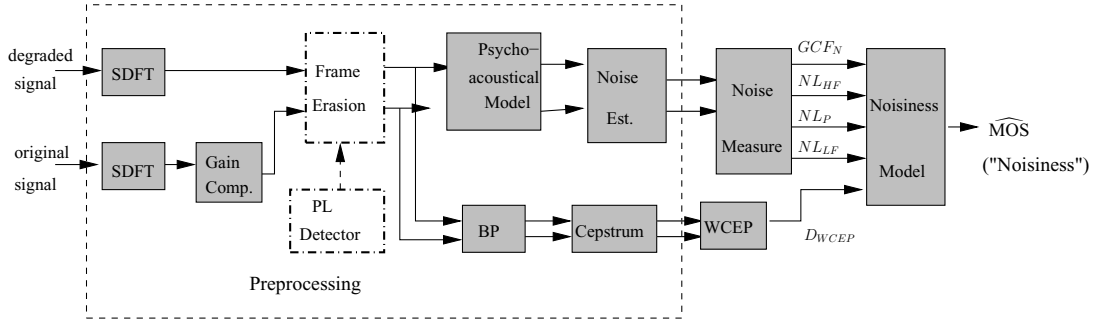

Figure 2: *Signal processing for "Noisiness" prediction.*

transfer function $H(\mu)$ of a transmission system is then estimated by a comparison of the clean and the degraded spectra. Finally, the system gain is equalized by

$$X'(\mu, i) = X(\mu, i) \cdot H(\mu). \tag{1}$$

3.1.2 Frame Erasion due to Packet-loss

If possible, the frames corrupted by packet-loss should be erased from consideration.

3.1.3 Critical-band filters

In order to obtain a spectral representation that better corresponds to the human auditory system, *critical-band filters* [8] were applied. In our case, $X'(\mu, i)$ and $Y(\mu, i)$ are filtered by 34 Gaussian-shaped critical-band filters presented in Fig. 3. The filter-band spectral represen-

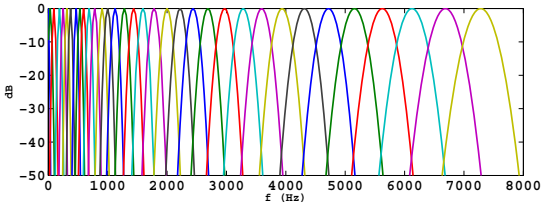

Figure 3: *34 critical bands used in the preprocessing.*

tation of $X'(\mu, i)$ can be obtained by:

$$\Phi_x(c, i) = \frac{\sum_{\mu=0}^{M-1} |X'(\mu, i)|^2 F(\mu, c)}{\sum_{\mu=0}^{M-1} F(\mu, c)} \tag{2}$$

where $F(\mu, c)$ is the μ-th coefficient of the c-th Gaussian-shaped filter. The spectral representation of the degraded signal $\Phi_y(c, i)$ can be determined analogously.

3.1.4 Noise Estimation

The noise-spectrum estimate $\Phi_n(c, i)$ is calculated as the difference between the degraded- and clean-signal spectral represntations in speech pauses:

$$\Phi_n(c, i) = \begin{cases} \max\{\Phi_y(c, i) - \Phi_x(c, i), \sigma\} & \text{in speech pauses} \\ \text{missing} & \text{else} \end{cases} \tag{3}$$

Here, σ is a small positive value.
The noise estimate for each band is then given by:

$$NL(c) = \text{median}\{10\log_{10}|\Phi_n(c, i)|\}. \tag{4}$$

In the final step, the A-curve $A(c)$ according to [9] in dB is applied to $\Phi_n(c, i)$:

$$NL_A(c) = NL(c)A(c). \tag{5}$$

3.1.5 Bandwidth Limitation

Parallel to the critical-band filtering and noise estimation, both $X'(\mu, i)$ and $Y(\mu, i)$ are filtered by the same rectangular bandpass filter so that the energy beyond the speech-relevant frequency range is totally suppressed. This step is necessary for the weighted cepstral distance measure so that it is only influenced by the "in-band" noise.

3.1.6 Cepstrum

The cepstrum is then calculated based on the bandpass-filtered spectra:

1. Inverse transform each frame into time domain,
2. Use Levinson-Durbin recursion to calculate the linear prediction coefficients $a(k, i)$ of order 8, where k indicates k-th coefficient.
3. Extract cepstral coefficients according to [10]

$$c(k, i) = a(k, i) + \sum_{l=1}^{k-1} \frac{l}{k} c(l, i) a(k - l, i). \tag{6}$$

We use $c_x(k, i)$ and $c_y(k, i)$ to denote the cepstral coefficients for the clean and the degraded signals.

3.2 Parameter measures

3.2.1 Determination of n_p, n_{lf} and n_{hf}

The parameters n_p, n_{lf} and n_{hf} represent the mean energies of all bands, the first 25 bands, and the last 9 bands, respectively.

$$n_p = 10\log_{10}\left(\frac{1}{34}\sum_{c=1}^{34} 10^{NL_A(c)/10}\right), \tag{7}$$

$$n_{lf} = 10\log_{10}\left(\frac{1}{25}\sum_{c=1}^{25} 10^{NL_A(c)/10}\right),$$

$$n_{hf} = 10\log_{10}\left(\frac{1}{9}\sum_{c=26}^{34} 10^{NL_A(c)/10}\right).$$

3.2.2 Determination of $f_{c,n}$

In order to obtain $f_{c,n}$, the amplitude noise spectrum is shifted so that its maximum is fixed as a positive value $ST = 10$ dB and its negative values are set to zero in the first step:

$$NL_{ST}(c) = \max\{NL(c) - NL\max + ST, 0\}. \quad (8)$$

$f_{c,n}$ is then determined by

$$f_{c,n} = \frac{\sum_{c=1}^{34} NL_{ST}(c) \cdot c}{\sum_{c=1}^{34} NL_{ST}(c)}. \quad (9)$$

3.2.3 Determination of d_{cep}

The cepstral distance is a well-known objective measure for the speech distortion [6]. Here we used a modified cepstral distance which is weighted by the speech energy:

$$d_{cep} = \frac{\sum_{i=0}^{L} w(i)\big(C\sqrt{\sum_{k=0}^{7}(c_x(k,i) - c_y(k,i))^2} - 1\big)}{\sum_{i=0}^{L} w(i)} \quad (10)$$

where $C = 10\frac{\sqrt{2}}{\log(10)}$ is a normalization constant according to [10] and $w(i)$ is the weighting factor determined by the signal energy of the i-th frame:

$$w(i) = \max\big\{20\log_{10}\frac{1}{M}\sum_{\mu=0}^{M-1}|X_B(\mu,i)| - (-30), 0\big\}. \quad (11)$$

3.2.4 Parameter Boundaries

As a last step, the parameter domains are restricted in order to take interactions between parameters and saturation effects into account:

$$\overline{n_{lf}} = \begin{cases} 0 & \text{if } n_{lf} > 0 \\ n_{lf} & \text{if } -35 \le n_{lf} \le 0 \\ -35 & \text{if } n_{lf} > -35 \end{cases} \quad (12)$$

$$\overline{d_{cep}} = \begin{cases} 3.5 & \text{if } d_{cep} > 3.5 \\ d_{cep} & \text{if } 1.5 \le d_{cep} \le 3.5 \\ 1.5 & \text{else} \end{cases} \quad (13)$$

$$\overline{n_p} = \begin{cases} -15 & \text{if } n_p > -15 \\ n_p & \text{if } -50 \le n_p \le -15 \\ -50 & \text{if } n_p > -50 \end{cases} \quad (14)$$

$$\overline{n_{hf}} = \begin{cases} 0 & \text{if } n_{hf} > 0 \& f_{c,n} > 22 \\ n_{hf} & \text{if } -40 \le n_{hf} \le 0 \& f_{c,n} > 22 \\ -40 & \text{else} \end{cases} \quad (15)$$

4 Dimension Estimator

Two steps are undertaken to train the dimension estimator for "noisiness". Firstly, estimators for the sub-dimensions are trained based on the SD scores from the auditory test and the above measured parameters. Then, the "Noisiness" estimator is trained on the basis the SD scores and the MOS values from the auditory tests.

Combining the above two models leads to a two-step prediction model for the overall quality, which indicates "noisiness" in our context.

As mentioned before, the positions of the BG noise in the 3D space are unstable and thus hard to predict. Hence the training of all the following prediction models have been done without these stimuli. However, as we will see later, the resulting model can also predict the overall quality of these stimuli quite well.

4.1 Sub-dimension Estimators

The SD prediction models are trained using the SD scores and the modified measures of one female speaker.

$$\widehat{sd_1} = \begin{cases} -5.62 + 3.84\overline{d_{cep}} - 0.51\overline{d_{cep}}^2, & n_{lf} \le -35 \\ 2.25 + 0.13\overline{n_{lf}} + 0.0011\overline{n_{lf}}^2, & \text{else.} \end{cases} \quad (16)$$

$$\widehat{sd_2} = 0.59 - 0.074\overline{n_p} - 0.0024\overline{n_p}^2 \quad (17)$$

$$\widehat{sd_3} = 1.30 - 0.036\overline{n_{hf}} - 0.0014\overline{n_{hf}}^2 \quad (18)$$

Fig. 4 shows the curve fitting results of the above formula to the signal measures. Table 1 summarizes the prediction performance for the training stimuli.

	without BG noise			with BG noise		
	N	ρ	RMSE	N	ρ	RMSE
sd_1	60	0.97	0.26	69	0.95	0.36
sd_2	60	0.97	0.26	69	0.82	0.59
sd_3	60	0.93	0.70	69	0.49	0.74

Table 1: *Performance of SD predictors for the training stimuli of one female speaker. N: sample size.*

4.2 Prediction of "Noisiness"

The prediction model for MOS is trained on the basis the SD scores and MOS values from the auditroy tests:

$$\widehat{MOS} = 2.660 - 0.531sd_1 - 0.2873sd_1^2 \quad (19)$$
$$-0.440sd_2 - 0.255sd_2^2 + 0.284sd_1sd_2 - 0.491sd_3$$

With this model, MOS values can be predicted by the SD scores with a correlation of $\rho = 0.97$ without BG noise and $\rho = 0.85$ with BG noise.

4.3 Prediction Performance

Substituting sd_1, sd_2 and sd_3 by $\widehat{sd_1}$, $\widehat{sd_2}$ and $\widehat{sd_3}$ in(19) and combining (16)–(19) results in an overall-prediction model.

Besides the above training stimuli, the stimuli from one male speaker are used as the test stimuli. Table 2 and Fig. 5 show the prediction performance of this two-step prediction model. Although the stimuli with BG noise can not be captured by all the prediction models, they can be predicted by the overall prediction model quite well.

	without BG noise			with BG noise		
	N	ρ	RMSE	N	ρ	RMSE
training	36	0.96	0.31	42	0.94	0.33
test	36	0.95	0.33	42	0.94	0.37

Table 2: *Performance of two-step MOS predictor. N: sample size.*

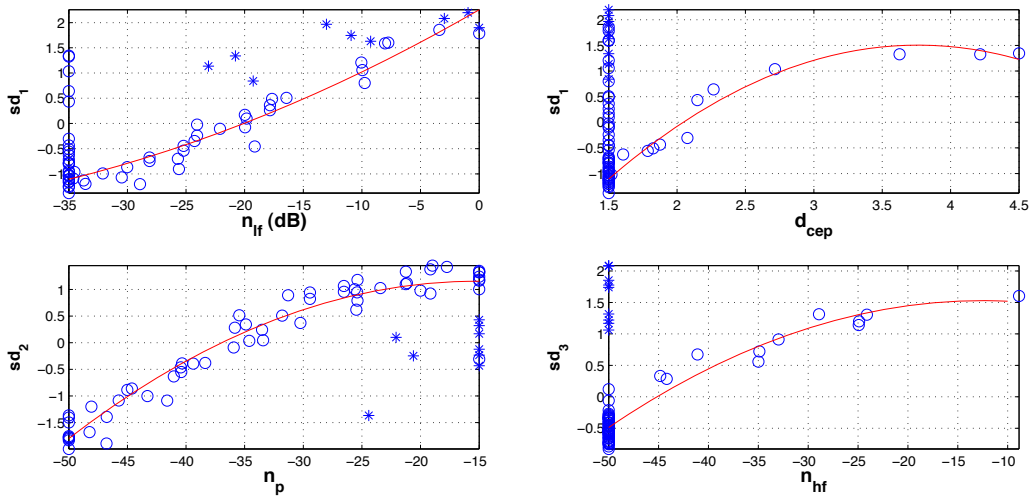

Figure 4: *Curve fitting of signal measures on target sub-dimensions. Star: BG noise; circle: the others.*

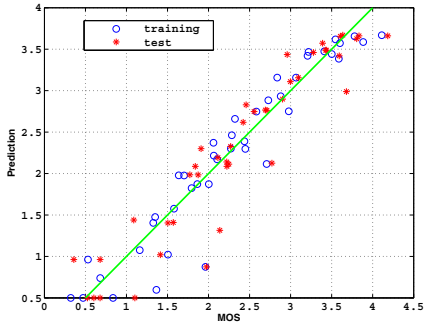

Figure 5: *Performance of two-step MOS predictor.*

5 Conclusions and discussion

In this paper, we have proposed an estimation model for the speech-quality dimension "Noisiness" which is based on so-called sub-dimensions. This model implies that "noisiness" can be divided into three meaningful sub-dimensions, "speech contamination", "(perceptual) additive noise level", and "noise coloration". Each of these sub-dimension can be predicted by combining the instrumental measures n_{lf}, n_{hf}, n_p, $f_{c,n}$ and d_{cep}, and the integral speech quality in the context of noise, can be predicted based on these predicted sub-dimensions.

Although the overall prediction model estimates the integral quality with a high correlation ($\rho \geq 0.94$) for all the stimuli, we have found that a group of stimuli with realistic background noise can not be captured in all prediction models. The good performance of the overall prediction model seems to support our assumption that this deviance may stem from the uncertainty felt by the test persons regarding this kind of "informative" and "complex" noise, however the exact reason of their deviance from the rest is still unknown.

References

[1] 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", *ITU-T*, Geneva, 2001.

[2] M. Wältermann, K. Scholz, A. Raake, U. Heute and S. Möller: "Underlying quality dimensions of modern telephone connections", *Int. Conf. on Spoken Language Processing*, page 210, Pittsburgh, USA, September 2006.

[3] I. Borg and P. Groenen: *Modern Multidimensional Scaling – Theory and Applications*. Springer Series in Statistics, USA–New York NY, 2 edition, 2005.

[4] L. Tsogo, M. H. Masson, and A. Bardot: Multidimensional scaling methods for many-object sets: A review. *Multivariate Behavioral Research*, 35(3):307–319, 2000.

[5] S. Möller: *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publishers, USA–Boston MA, 2000.

[6] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements: *Objective Measures of Speech Quality*, Prentice Hall Advanced Reference Series, Englewood Cliffs, NJ, 1988, ISBN: 0-13-629056-6.

[7] K. Scholz, C. Kühnel, M. Wältermann, U. Heute and S. Möller: "Assessment of the Speech-Quality Dimension "Noisiness" for the Instrumental Estimation and Analysis of Telephone-Band Speech Quality", submitted to *Int. Conf. on Spoken Language Processing*, 2008.

[8] E. Zwicker;H. Fastl: *Psychoacoustics: Facts and Models*, Springer, Berlin, 1999.

[9] ITU-R Rec. BS.468: *Measurement of Audio-Frequency Noise Voltage Level in Sound Broadcasting*, ITU, 1990.

[10] P. Vary, R. Martin: *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, John Wiley & Sons LTD, 2006.