Can latent speech quality dimensions be quantified directly?

Marcel Wältermann, Alexander Raake, and Sebastian Möller

Deutsche Telekom Laboratories, TU Berlin, Germany

{marcel.waeltermann, alexander.raake, sebastian.moeller}@telekom.de

Abstract

Previous studies revealed that the quality of transmitted speech can well be described by three perceptual dimensions: "discontinuity", "noisiness", and "coloration". In this paper, a method is presented for quantifying these dimensions directly in subjective tests. Two experiments were conducted according to this method with large sets of test conditions. A detailed description of the test procedure and a thorough analysis of the results is presented. The proposed method is reliable and produces meaningful and orthogonal results for diverse stimuli sets, in particular for stimuli containing multi-dimensional degradations. The method might form the basis for collecting the data for future diagnostic speech quality estimators.

Index Terms: Speech quality, modeling, feature decomposition

1. Introduction

The transmission of human speech signals through any communication system or network leads to a modification of the original physical signal emitted from the mouth of the speaker. In most of the cases, this causes a deterioration of the speech quality as perceived by the listener at the other end of the communication channel. It is of interest to system and network designers to determine the amount and the type of such deterioration. A valid and reliable way for doing this is to conduct listening or conversation experiments with a number of human listeners or interlocutors and letting them judge the quality by means of an appropriate scale. For speech quality experiments, the requirements and procedures are described in ITU-T Rec. P.800.

The quality ratings obtained in subjective experiments are usually of integral nature. That is, they are the result of an integrative reflection process of the listener [1]. The obtained ratings do not necessarily allow conclusions to be drawn with regard to the perceptual aspects that cause a particular quality rating. For instance, both the perceptual effect of a certain level of noise and a certain packet loss rate could lead to the same quality rating, whereas the roots for it remain hidden. Diagnostic quality information, however, might be of additional value, especially if it is of perceptual nature and thus reflecting the actual importance of physically provoked distortions.

In general, such features can be represented by attributes describing the perceptual event in a semantically meaningful way. These attributes can be used as scale labels in order to quantify the perceptual features in subjective tests. Usually, analytic tests published in the relevant literature (e.g., [2][3]) are based on a large set of these descriptive attributes being rated by listeners. In order to obtain a more parsimonious and thus clearer picture of the collected data, procedures like Principal Component Analysis help to summarize correlating attributes to a reduced number of common orthogonal components, so-called latent dimensions.

In this paper, a new method is presented for assessing such underlying dimensions *directly* by listeners, emanating from the knowledge of demonstrably orthogonal dimensions for transmitted speech as revealed by previous studies. The benefit of such a procedure is the reduced experimental effort due to the reduced number of necessary scales and thus the number of judgments per stimulus. Hence, diagnostic instrumental speech quality estimation models [4] that rely on subjective databases can more effectively be trained. Moreover, as it was shown in [5], the orthogonal features describe integral quality of transmitted speech in a nearly complete way. Thus, the diagnostic estimates can be employed to predict integral speech quality.

Previous exploratory research aiming at the revelation of the latent dimensions is summarized in Section 2. In Section 3, the method is presented in detail, including the rating scale design and the test procedure. The developed method was applied in two listening experiments with a large number of conditions. Section 4 contains the description and the analysis of these experiments. Evidence is given for the meaningfulness, orthogonality, and test-retest reliability. Finally, conclusions and a discussion about possible extensions to the method is provided in Section 5. Compared to [5], the present paper focusses on an extensive description of the method and two experiments, as well as an in-depth analysis of the data.

2. Previous Research

In previous studies [6][7][8], extensive auditory tests were carried out aiming at revealing the underlying dimensions of speech quality. Two different paradigms were followed: a) Similarity scaling of pairwise presented stimuli, and b) attribute-scaling in the fashion of Semantic Differentials where a larger set of (not necessarily orthogonal) attributes are judged for each stimulus. The test conditions included electro-acoustic properties of user terminals, send-side room acoustics, narrowband (NB, 300-3400 Hz) and wideband (WB, 50-7000 Hz) codecs, speech enhancement algorithms, and variation of behavior of the underlying network (e.g., packet loss). The set of attributes for the Semantic Differential was derived from additional experiments and extended using those attributes found in the relevant literature (e.g., [2]).

The experimental data obtained by these tests can be represented in spaces of low dimensionality by the methods of a) Multidimensional Scaling, and b), e.g., Principal Component Analysis. The spaces' axes reflect the perceptual dimensions. The interpretation of the underlying axes led to the following dimension labels (degradation types associated with the respective dimension are given in brackets):

- *Discontinuity* (packet loss, silence insertion, time-varying effect of signal-correlated noise, time-varying codec non-linearities, musical noise),
- *noisiness* (signal-correlated noise, additive circuit and background noise), and

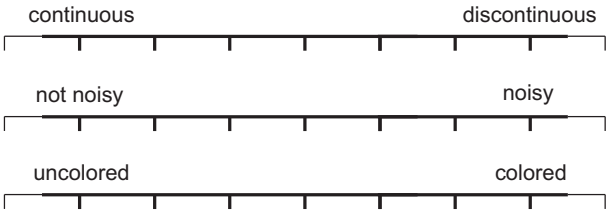

Figure 1: *Dimension scales.*

- *coloration* (linear distortions due to bandpass filtering and room acoustics).

3. Direct Assessment of the Latent Dimensions

3.1. Rating Scales

In order to support the applicability of the dimensions identified in the exploratory experiments described in Section 2, a new assessment method is presented here which provides a means for quantifying the above-mentioned latent speech quality dimensions directly by means of three descriptive scales, each dedicated to one dimension. This way, time-consuming Multidimensional Scaling and/or Semantic Differential can potentially be avoided. The three scales were designed for measuring the three dimensions "discontinuity", "noisiness", and "coloration", respectively. The poles of the scales are labeled with the antonym attributes "continuous – discontinuous" ("discontinuity" dimension), "not noisy – noisy" ("noisiness" dimension), and "uncolored – colored" ("coloration" dimension).[1] That way, separate scores for the perceptual dimensions present in degraded speech can be obtained.

The graphical layout is depicted in Figure 1. It is similar to the scale layout recommended in ITU-T Rec. P.851. Since the labels on the left ends of the scales describe zero impairment in the respective dimension, whereas the labels on the right ends describe maximum impairment, the scales can be considered as being unipolar scales.

3.2. Test Procedure

In the experiments described in Section 4, each speech sample is assessed taking into account the three scales depicted in Figure 1. Prior to the scaling, the meaning of the scales is explained to the listeners during dedicated training phases. Therefore, a detailed written description of the three dimension scales is given to the subjects. The instructions start off explaining that in this part of the experiment, the features or characteristics of speech samples are supposed to be judged (i.e., not the quality), and that this evaluation is done by means of three scales. Each scale is labeled with an attribute at each end that describes the characteristic to be judged upon. Each scale and its usage are separately described in detail, using synonyms to the scale attributes as an aid. According to a Principal Component Analysis of Semantic Differential data (cf. Section 2), the synonyms chosen here correspond to those attributes which are very highly correlated with the principal components reflecting the perceptual dimensions (see, e.g. [6]).

Furthermore, exemplary samples for each scale are presented which are distorted in (mainly) one dimension:

[1]Translation from the German wordings "kontinuierlich – diskontinuierlich", "unrauschhaftig – rauschhaftig", and "klanglich unverfärbt – klanglich verfärbt".

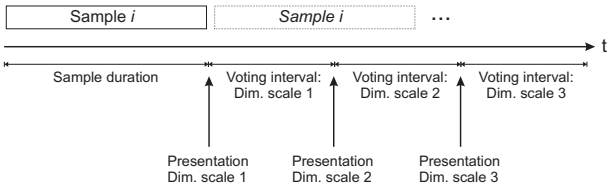

Figure 2: *Sample and scale presentation and rating.*

- "Discontinuity": G.722.2 (23.05 kbit/s) with 2% and 8% packet loss
- "Noisiness": WB PCM (50-7000 Hz) with stationary car noise at send side, 70 dB(A)
- "Coloration": WB and NB bandpass filters of different shape

The acoustic presentation is done together with the describing synonyms. The participants can listen to the samples until they confirm that they understand the meaning of the scales. The understanding is supported by presenting an undistorted sample (direct WB), stating that this particular sample is completely "not noisy", "continuous", and "uncolored".

Preceding trials help the listeners familiarizing with the practical usage of the scales and the range of degradations to be expected. Therefore, as in the overall quality test, several samples differing in quality and character of the degradation were rated in a brief dedicated training session. In the dimension assessment experiment, the scales are presented separately, i.e. consecutively for each stimulus. In prior to the registering of the ratings, the listeners are asked to listen to the entire speech sample. During one trial, they can optionally repeat the playback. The rating scheme for one sample is depicted in Figure 2. The samples are presented in randomized order. For each participant, the order of the scales is permutated. The order is held constant per participant in order to avoid confusion of the scales. Depending on the number of test stimuli, the dimension assessment needs to be subdivided into different sessions in order not to produce listener fatigue.

4. Application and Analysis of the Method

4.1. Test Conditions

Two auditory experiments were carried out according to the protocol described in Section 3. Each experiment was designed to test the performance of the scales for particular kinds of degradations. Experiment 1 was intended to provide insight into

- the perceptual composition of codecs and codec tandems with regard to the three dimension scales, and
- the perception of the coloration of codecs and "equivalent linear-only distortions".

A total of 66 processing chains were considered in Experiment 1: 8 NB and 7 WB codecs at various bitrates, for instance the G.711, the G.726, the G.722, and the G.722.2, one "clean" WB PCM condition, 24 codec tandems, and 14 conditions including noise, packet loss, and bandpass filters. For 12 conditions, no codecs were applied to the source speech files. In fact, these conditions contain the "linear portion" of the overall distortion of some of the codec or codec tandem conditions. This was done by estimating the magnitude of the transfer function of the corresponding codec condition, designing an FIR-Filter of equal shape, and applying that filter to the source speech material (cf. [9]). The aim here was to cross-check whether the

"coloration" assessment for regular codecs is equal to the rating of the corresponding purely linear distorted conditions, and whether the "discontinuity" and "noisiness" scales are insensitive to linear distortion.

In contrast, Experiment 2 focuses on

- the perceptual composition of packet loss,
- the perceptual composition of realistic background noise at send side, and
- the perceptual composition of realistic background noise at send side with simultaneous packet loss.

In this Experiment, 76 processing chains were included: 6 codecs (NB and WB), 6 NB and WB MNRU conditions (Modulated Noise Reference Unit according to ITU-T Rec. P.810; modified for WB), 20 conditions with different codecs/packet loss concealment (PLC) schemes and packet loss percentages Ppl (uniform loss pattern, $Ppl \in \{1, 2, 4, 8\}$ %), 29 conditions with send-side background noise of different type (car, pub, cafeteria) and level Ps ($Ps \in \{45, 55, 70\}$ dB(A)), 12 conditions with combinations of background noise and packet loss, and 3 bandpass filters. Among these Experiment 2 conditions, there are 12 reference conditions already included in Experiment 1. These reference conditions include the clean WB PCM, different codecs at different bitrates, bandpass filters, noise of different levels and different rates of packet loss. They were selected in such a manner that they are as evenly spread over each of the three dimension scales as possible. These conditions serve to check the reliability of the proposed method (see Section 4.3.1).

The speech source material was sampled at 48 kHz. Each sample had a duration of 10s on average. Different German sentences uttered by one male and one female speaker were considered. Other experimental conditions, e.g. with regard to the listening environment, complied with ITU-T Rec. P.800. The presentation of the stimulus material was done diotically using headphones. The listening level was set to 73 dB SPL.

4.2. Listeners

Two independent groups of native German listeners were recruited which mostly consisted of students from the local university: 20 listeners (10 f, 10 m) for Experiment 1, 24 listeners (12 f, 12 m) for Experiment 2. They were aged between 20 and 33 (average age: 27.3) in Experiment 1, and between 20 and 46 (average age: 28.7) in Experiment 2. None of them reported any known loss of hearing and they were paid for their participation.

4.3. Results

4.3.1. General Characteristics of the Data

In the following, some key characteristics of the raw scale data $S_{dim} \in [0; 1]$, with $dim \in \{\text{dis}, \text{noi}, \text{col}\}$, common to both experiments are discussed.[2]

The means of the standard deviations $\varnothing std_{dim}$ were calculated on a per-file basis. They amount to $\varnothing std_{\text{dis},1} = 0.195$ (Exp. 1) / $\varnothing std_{\text{dis},2} = 0.211$ (Exp. 2), $\varnothing std_{\text{noi},1} = 0.177$ (Exp. 1) / $\varnothing std_{\text{noi},2} = 0.232$ (Exp. 2), and $\varnothing std_{\text{col},1} = 0.202$ (Exp. 1) / $\varnothing std_{\text{col},2} = 0.191$ (Exp. 2) for the "discontinuity" scale, the "noisiness" scale, and the "coloration" scale, respectively. These values lie well within the range of standard deviations obtained on ACR scales (see, e.g., [10], p.151).

[2]Note that S_{dim} corresponds to S'_{dim} in [5]. The hyphen is omitted here for simplicity.

The scales were used in an orthogonal way; there is a very low correlation $r_{dim1,dim2}$, with $dim1 \neq dim2$ between the ratings on every two scales. For Experiment 1 (Experiment 2), the correlation coefficients amount to $r_{\text{dis,noi}} = 0.040$ ($r_{\text{dis,noi}} = -0.015$), $r_{\text{dis,col}} = 0.046$ ($r_{\text{dis,col}} = 0.061$), and $r_{\text{noi,col}} = 0.251$ ($r_{\text{noi,col}} = 0.086$).

In order to get more insight into the nature of the per-condition scores, the per-condition distributions were investigated before analyzing the mean scores $\bar{S}_{dim}$. Due to the limited space of this paper, we describe the characteristics of the score distributions only qualitatively here. One-sample Kolmogorov-Smirnov tests only occasionally indicate significant deviation of the dimension scale data from normality. Systematic exceptions consist of high (low) scores S_{dim}: The distributions show relatively high negative (positive) skewness and positive kurtosis, together with a relatively low standard deviation for conditions which are not degraded (highly degraded) in the dimension captured by the scale. Apparently, the listeners judge them as being optimal (not optimal) and use the respective scale label more or less consistently as an anchor. For degradations which are rated on the continuum between the anchor labels, the magnitude of the skewness typically decreases with a simultaneous increase of the standard deviation, reflecting an uncertainty between the listeners. For some types of degradations, however, bimodal distribution emerge, representing groups of listeners with two contradictory "points-of-view" how the degradation should be judged. This effect manifests itself in a negative Kurtosis (and a high standard deviation). Such observations will be discussed in detail in the following sections. Although the prerequisite of normality of the data is not met for all conditions, single univariate mixed-model ANOVAs were separately applied to the data obtained from the "discontinuity" scale, the "noisiness" scale, and the "coloration" scale, respectively. This can be justified since ANOVA is relatively robust with regard to violations of this and other (variance homogeneity etc.) formal requirements (see, e.g., [11]).

In the mixed-model ANOVAs, the factor *subject* was included as a random variable, whereas the remaining experimental factors were fixed: *Speaker*, *codbit* (implicit combination of codec and bitrate; codec tandems are also subsumed here), *pl* (packet loss), and *filter*, referring to the bandpass filters used. Moreover, the factor *noise* was included in the Experiment 1 analysis, whereas *ntype* (noise type) and *nlevel* (noise level) were included in Experiment 2 instead. Note that the MNRU conditions are subsumed under the *codbit* factor here. Hence, *ntype* and *nlevel* only reflect additive noise, and not signal-correlated noise. The main effects and 2-way interactions (where possible) were tested.

The ANOVA's for the dimension scales reveal that the *subject* and *speaker* factors often turn out to interact significantly with other experimental factors. For instance, the interactions *codbit*subject* and *speaker*subject* appear consistently to be significant over all scales and tests, as it is almost always also the case for *speaker*codbit*. Apparently, there are individual differences in scale usage for different speakers and codecs, and a speaker-dependent codec-judgment. The subject-dependency is, however, not caused by the test method itself which is presented and analyzed here. The listener groups of both tests also participated in standard ACR tests according to ITU-T Rec. P.800 and rated the overall quality of both sets of conditions. Mixed-model ANOVAs reveal similar significant interaction effects with the factors *subject* and *speaker* (data not reproduced here), thus being seemingly not avoidable. Other interactions such as *pl*subject* ("discontinuity" scale), *noise/ntype*subject*

("noisiness" scale), and *filter*subject* ("coloration" scale) can also be observed for the MOS data. Moreover, the speaker and subject dependencies are generally weaker than the experimental factors actually of interest and thus of minor practical relevance. The partial Eta-squared η_{p}^2 reflects the proportion of total variation of one factor, excluding other factors. In general, the influence of *subject* and *speaker* main effects and interactions for the MOS scale and the three dimension scales is quite comparable. Thus, for practical application, averaging the dimension scale values S_{dim} over subjects and speakers for every scale seems to be justified. In this paper, we will consider the arithmetic mean $\bar{S}_{dim}$ in the following.

As a final remark, the 12 common conditions included in Experiments 1 and 2 allow to provide some insight into the test-retest reliability of the subjective method presented here. The correlation coefficients r and root mean square errors ($RMSE$) between the mean raw scale ratings $\bar{S}_{dim}$ of Experiments 1 and 2 amount to $r_{\mathrm{dis1,dis2}} = 0.96$, ($RMSE_{\mathrm{dis1,dis2}} = 0.10$), $r_{\mathrm{noi1,noi2}} = 0.98$, ($RMSE_{\mathrm{noi1,noi2}} = 0.08$), and $r_{\mathrm{col1,col2}} = 0.98$, ($RMSE_{\mathrm{col1,col2}} = 0.09$). The close-to-linear agreement between the absolute scale values of the two tests provide evidence that the method is reliable to a high degree.

4.3.2. Experiment 1

The ANOVA results of Experiment 1 can be found in Table 1 (all subject and speaker effects omitted, cf. discussion in Section 4.3.1; only significant effects at the 1%-level, i.e. $p < 0.01$).

Table 1: *ANOVA Experiment 1 (relevant effects only).*

Scale	Source	df_{n}	df_{d}	F	p	η_{p}^2
dis	**codbit**	39	741	16.1	.000	.458
	pl	4	76	112.0	.000	.855
noi	**codbit**	39	741	30.1	.000	.613
	noise	4	76	159.8	.000	.894
	filter	14	276.9	3.6	.000	.155
	codbit*noise	2	1275	44.8	.000	.066
col	**codbit**	39	741	32.8	.000	.633
	filter	14	275.1	36.0	.000	.647

From Table 1, it can be concluded that the factors *codbit* and *pl* have significant influence on S_{dis}. While the latter effect will systematically be investigated in the next section for different rates of packet loss, the "discontinuity" perception with increasing bitrate (dependent on the codec family) can be seen from the black bars in Figure 3.

Codec modes of high bitrate are perceived as continuous (low $\bar{S}_{\mathrm{dis}}$ values), comparable to the direct channel (WB PCM). For instance, the G.711 condition is perceived as equally continuous as the WB PCM is. Signal-correlated, and thus time-varying noise originating from ADPCM codecs (G.722, G.726) leads to an increase in "discontinuity" perception, resulting in monotonically increasing values $\bar{S}_{\mathrm{dis}}$ with decreasing bitrate. In [3], the perceptual effect from non-linear distortion of low-bitrate (NB) codecs has been described with the term "bubbling". It appears that this effect is captured by $\bar{S}_{\mathrm{dis}}$, as the "discontinuity" perception systematically increases with decreasing bitrate of hybrid codecs (e.g., the G.722.2). This effect can also be observed comparing the $\bar{S}_{\mathrm{dis}}$ values of the G.711 with G.729A and GSM-EFR.

The S_{noi} ratings are mostly influenced by the main factors *noise* and *codbit*, as expected (see Table 1). The significant influence of the factor *filter* is due to the low $\bar{S}_{\mathrm{noi}}$ value

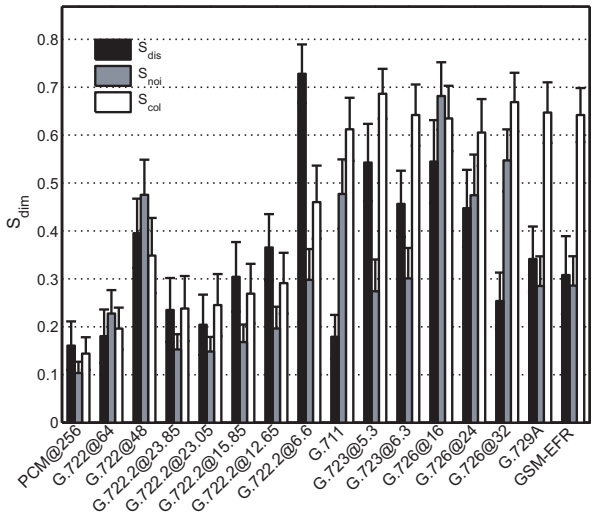

Figure 3: $\bar{S}_{dim}$ *values with 95% CIs for various codecs.*

for one of the filtered NB conditions in Experiment 1, which cannot be explained in a plausible way. The relation between the factor *noise* (noise type and level) will systematically be investigated in the course of the analysis of Experiment 2 in the next section. The gray bars in Figure 3 illustrate the $\bar{S}_{\mathrm{noi}}$ ratings for different codecs and bitrates. It can be observed that the $\bar{S}_{\mathrm{noi}}$ increases for ADPCM codecs (G.722, G.726) with decreasing bitrate, with the exception of the 24 kbit/s and 32 kbit/s modes of the G.726 which cannot plausibly be explained. Also, $\bar{S}_{\mathrm{noi}}$ increases with decreasing bitrate of hybrid codecs (e.g., the G.722.2). Due to the noticeable noise-floor of the G.711 condition, $\bar{S}_{\mathrm{noi}}$ is relatively high as compared to other NB codecs such as the G.723, the G.729A and the GSM-EFR. The significant interaction effect *codbit*noise* stems from the different noise perception in NB and WB, which will be investigated further in the course of the next experiment.

Table 1 reveals that the S_{col} data significantly depend on the main effects *codbit* (inherent bandwidth restrictions of NB and WB codecs) and *filter*. The white bars in Figure 3 graphically illustrate $\bar{S}_{\mathrm{col}}$ values for different codecs and bitrates. For WB codecs, the values $\bar{S}_{\mathrm{col}}$ increase with decreasing bitrate. As this is perceptively plausible for the G.722.2 codec, the relatively high value for G.722 (48 kbit/s) as compared to the 64 kbit/s version is astonishing for an expert, since the audio bandwidth is more or less equal for either bitrates. An explanation might be that the listeners interpret more or less severe signal-correlated noise as "coloration". This is, however, not true in the NB case, where the $\bar{S}_{\mathrm{col}}$ values for the G.726 are approximately equal. This phenomenon will further be discussed in the next section, where similar effects can be observed for MNRU conditions. For NB codecs, the ratings $\bar{S}_{\mathrm{col}}$ are approximately constant. Thus, the audio bandwidth is well captured by the "coloration" scale.

Figure 4 provides some evidence that mainly the "linear portion" of codecs is captured by the "coloration" scale: The black bars represent $\bar{S}_{\mathrm{col}}$ values of the codec and tandem conditions, whereas the white bars depict the ratings of the purely linearly filtered conditions, where the filters correspond to the "transfer functions" of the real codecs (cf. [9]). Since the distance between corresponding bars is small, the concept of "coloration" was judged equally for both groups of conditions. The G.722 at 48 kbit/s seems to be an exception here. The "coloration" of the pure linear condition deviates very strongly from

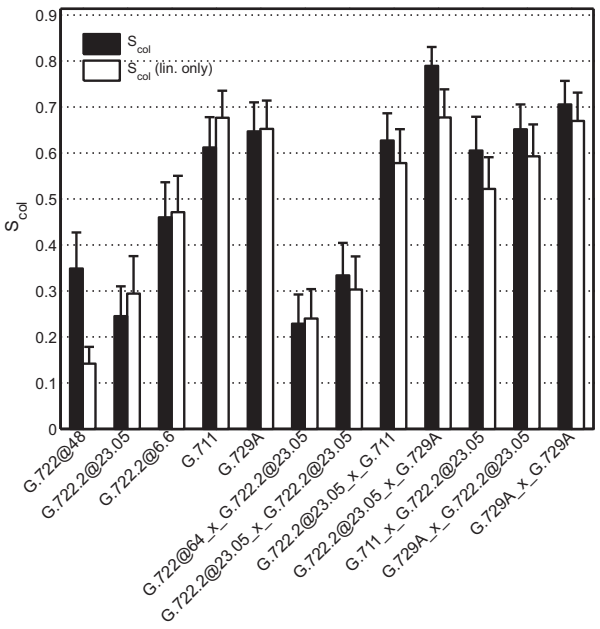

Figure 4: *Comparison of $\bar{S}_{\mathrm{col}}$ values with 95% CIs.*

that of the real codec condition. This is yet another indicator that the signal-correlated noise in the WB case obviously affects the values $\bar{S}_{\mathrm{col}}$.

4.3.3. Experiment 2

Analogous to the analysis of the first experiment, the ANOVA results of Experiment 2 can be found in Table 2 (all subject and speaker effects omitted, cf. discussion in Section 4.3.1; only significant effects at the 1%-level, i.e. $p < 0.01$).

Table 2: *ANOVA Experiment 2 (relevant effects only).*

Scale	Source	df_{n}	df_{d}	F	p	η_{p}^2
dis	**codbit**	11	520.0	26.9	.000	.363
	pl	4	202.2	63.0	.000	.555
	nlevel	2	182.2	4.8	.009	.050
	codbit*pl	16	2983	4.1	.000	.021
	pl*ntype	3	2983	4.9	.002	.005
noi	**codbit**	11	455.9	28.3	.000	.406
	ntype	2	65.2	8.0	.001	.198
	nlevel	2	97.9	115.6	.000	.702
	codbit*nlevel	8	2983	8.2	.000	.022
	ntype*nlevel	3	2983	34.1	.000	.033
col	**codbit**	11	414.9	126.9	.000	.771
	pl	4	389.3	4.7	.001	.046
	filter	2	58.2	21.9	.000	.429
	codbit*pl	16	2983	2.1	.008	.011

With regard to the values S_{dis}, the main factors *codbit* and *pl* are of highest significance (due to their highest F-ratios), see Table 2. The relatively high number of different codec-packet loss combinations (and thus the employed PLC techniques) obviously lead to the significant effect *codbit*pl*. In Figure 5, the relation between the packet loss rate (Ppl) and the values $\bar{S}_{\mathrm{dis}}$ is exemplarily shown for G.722 at 64 kbit/s and the G.711 codec. The individual curves reflect the codec-(PLC)-dependence. With higher packet loss rates, the codec-inherent discontinuity is superimposed with non-continuous artifacts from lost speech information and the specific PLC techniques. Thus, the curve shape differs slightly for the two codecs.

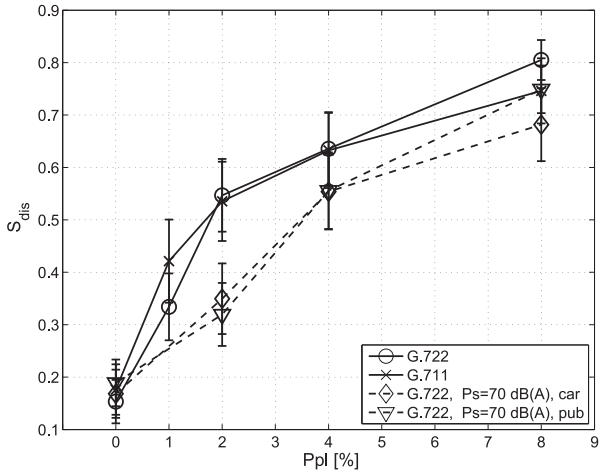

Figure 5: *$\bar{S}_{\mathrm{dis}}$ values with 95% CIs vs. packet loss percentage Ppl.*

The two remaining significant effects are provoked by the factor *nlevel*, and the interaction *pl*ntype*. Since *nlevel* is a main effect and there is no interaction with another inner-subject factor, it is likely that the noise level generally masks the "discontinuity" perception. In fact, one can see from Figure 5 that with a higher noise level of 70 dB(A), the values $\bar{S}_{\mathrm{dis}}$ are generally lower, independent from the type of noise. This is also true for discontinuous background noise as the cafeteria noise (not shown). For small and medium "discontinuities" caused by packet loss, it can be seen from Figure 5 that for higher noise levels, lower values $\bar{S}_{\mathrm{dis}}$ are obtained, indicating a masking of the discontinuous effect of packet loss by noise.

The $\bar{S}_{\mathrm{dis}}$ values for the MNRU conditions decrease monotonically for NB and WB, respectively, with decreasing signal-to-quantizing-noise ratio Q (not shown due to space restrictions). Thus, the fact that this distortion is timely varying due to its signal-correlatedness is captured by the "discontinuity" scale (cf. previous section). However, the standard deviations are remarkably high for these conditions. An inspection of the histograms (not included here) reveals that the ratings of very high Q-values do not follow a normal distribution, but are rather a bimodal one: There are two more or less clear "points-of-view" of the listeners on how MNRU conditions should be judged, either being more "continuous" (maybe the "noisiness" perception dominates here), or being more "discontinuous".

Besides the significant main effects *codbit* and *ntype*, *nlevel* is most important for S_{noi}. The (send side) noise level also significantly interacts with *codbit* and the type of noise, *ntype*. These relations are illustrated in Figure 6, again exemplarily for the G.711 and G.722 at 64 kbit/s codecs and car and pub type of noise. The different types of noises can plausibly be measured by the noisiness scale, even if the temporal structure of the noise is complex (cafeteria). For all codecs and noise types, the $\bar{S}_{\mathrm{noi}}$ values increase monotonically with increasing noise level Ps. The $\bar{S}_{\mathrm{noi}}$ values of the car and pub noises saturate for higher noise levels: The difference in $\bar{S}_{\mathrm{noi}}$ between medium and high noise levels does not result in much higher $\bar{S}_{\mathrm{noi}}$ values. For higher noise levels, the curves for the car noise, and the curves for the pub noise converge more or less in a single value $\bar{S}_{\mathrm{noi}}$, independent of the audio bandwidth or codec type. Generally, WB noise conditions achieve higher noise values than NB noise conditions for low noise levels (including the inherent codec noise). Overall, non-stationary, "informational" noise (pub, cafeteria) is rated lower in terms of $\bar{S}_{\mathrm{noi}}$ than stationary

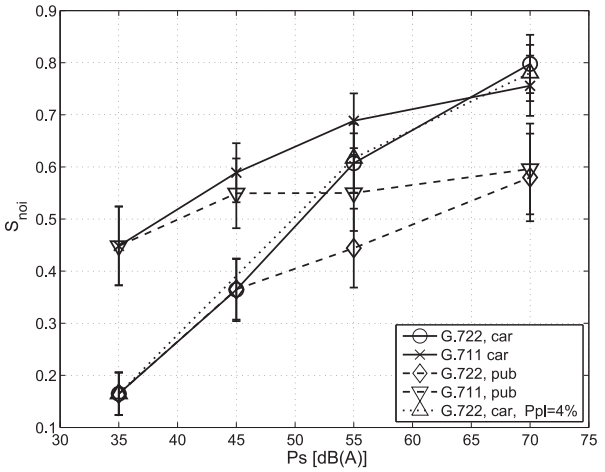

Figure 6: $\bar{S}_{\mathrm{noi}}$ *values with 95% CIs vs. background noise at send side.*

noise (car). Thus, it might be concluded that non-stationary noise with a more "airy" temporal structure is perceived as less noisy. Note that the temporal structure does not have any influence on "discontinuity" (see discussion above). It can also be seen from the dotted curve representing the G.722 at 64 kbit/s with $Ppl = 4\%$ (example) that for combined degradations (background noise at send side together with packet loss), there is no significant variation of $\bar{S}_{\mathrm{noi}}$ compared to the case of zero packet loss. In other words, there is no masking effect of "noisiness" perception coming from packet loss, at least if the G.722 (64 kbit/s) is employed. In the contrary, noise does influence the perception of "discontinuity", as it was seen above.

The standard deviation for the pub and cafeteria noise conditions, i.e. noise that is carrying a certain degree of information, are noticeably higher than for car noise conditions (not shown here). In fact, the distribution of higher levels of pub- and cafeteria noise (55 and 70 dB(A)) show distinctively negative kurtosis. Histograms of pub and cafeteria noises tend to be of a bimodal structure, with peaks at the scale extremes. Thus, there are apparently listeners with a different "point-of-view" with regard to the assessment of the "noisiness" of noise carrying information. An inclusion of non-stationary noises into the training (see Section 3.2) might help to overcome this problem.

Very high standard deviations can also be observed for the "noisiness" ratings of the MNRU conditions (not shown). Bimodal distributions of $\bar{S}_{\mathrm{noi}}$ can be observed here as well (the bimodal character of the histograms is particularly pronounced for high levels of Q). Like it is the case for the "discontinuity" of the MNRU conditions (see above), only a part of the listeners consider signal-correlated noise as actually being noisy, and might attribute it more to "discontinuity".

The results of the ANOVA with S_{col} as the dependent variable can be found in Table 2. As expected and as already seen from Experiment 1, the main factors *codbit* and *filter* are significant. In addition, the WB MNRU conditions provoke a slight increase of "coloration" perception, i.e. monotonically increasing $\bar{S}_{\mathrm{col}}$ values with decreasing Q-values (not shown). Although the standard deviation is relatively high, a slightly bimodal distribution of the listener ratings can only be observed for the strongest MNRU degradation in the WB case. Thus, the "coloration" perception of MNRU seems to be common consensus. For the NB MNRU, in the contrary, the "coloration" perception is constant, together with a low standard deviation. These observations are inline with those of Experiment 1. There, the G.722 (48 kbit/s) also had an impact on "coloration" perception. The ANOVA also suggests a significant effect of the packet loss rate (*pl*) and an interaction between the packet loss rate and the factor *codbit*. In fact, especially for WB codecs, the packet loss rate provokes certain degradation in terms of "coloration".

5. Conclusions

In this contribution, a method is presented which allows for subjectively rating single perceptual dimensions in test conditions which are distorted in one or multiple dimensions in a direct way. With the analysis of two experiments conducted following the proposed method, it has been shown that naive listeners are able to distinguish between the three dimensions "discontinuity", "noisiness", and "coloration" after a brief training phase. The direct rating of dimension impairments was done in a meaningful, orthogonal, and reliable way. Thus, we can answer in the affirmative to the question raised by the title of this paper. Moreover, this method can be considered as a basis for the development of diagnostic instrumental speech quality measures. However, a validation for different languages remains still to be done. The proposed method might be extended in order to handle degradations with non-optimal loudness. Although loudness degradations can be expressed as a unipolar impairment, the auditory measurement might only be possible with a bipolar scale.

6. References

[1] U. Jekosch, *Voice and Speech Quality Perception - Assessment and Evaluation*, ser. Signals and Communication Technology. D–Berlin: Springer, 2005.

[2] W. D. Voiers, "Diagnostic acceptability measure for speech communication," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'77)*, USA-Washington, 1977, pp. 204–207.

[3] V.-V. Mattila, *Perceptual Analysis of Speech Quality in Mobile Communications*. FIN–Tampere: Dissertation, Tampere University of Technology, 2001, vol. 340.

[4] U. Heute, S. Möller, A. Raake, K. Scholz, and M. Wältermann, "Integral and diagnostic speech-quality measurement: State of the art, problems, and new approaches," in *Proc. 4th European Congress on Acoustics (Forum Acusticum 2005)*, H-Budapest, 2005, pp. 1695–1700.

[5] M. Wältermann, A. Raake, and S. Möller, "Analytical assessment and distance modeling of speech transmission quality," in *Accepted for: 13th Int. Conf. on Spoken Language Processing (ICSLP 2010)*, JP–Makuhari, 2010.

[6] M. Wältermann, K. Scholz, A. Raake, U. Heute, and S. Möller, "Underlying quality dimensions of modern telephone connections," in *Proc. 9th Int. Conf. on Spoken Language Processing (ICSLP 2006)*, USA–Pittsburgh PA, 2006, pp. 2170–2173.

[7] M. Wältermann, A. Raake, and S. Möller, "Perceptual dimensions of wideband-transmitted speech," in *Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, D-Berlin, 4-6 September 2006, pp. 103–108.

[8] ——, "Quality dimensions of narrowband and wideband speech transmission," *submitted to acta acustica*.

[9] M. Wältermann and A. Raake, "Towards a new E-model impairment factor for linear distortion of narrowband and wideband speech transmission," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '08)*, USA-Las Vegas NV, 30 March - 4 April 2008, pp. 4817–4820.

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

[11] J. Bortz, *Statistik für Sozialwissenschaftler*. Berlin Heidelberg, 6. Aufl: Springer, 2005.