Underlying Quality Dimensions of Modern Telephone Connections

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

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

marcel.waeltermann@ruhr-uni-bochum.de

Abstract

It is the aim of the present paper to analyze the perceptual quality dimensions of modern telephone connections. Such connections differ from standard connections in their time-variant characteristics (e.g., due to Voice-over-IP transmission or due to noise reduction algorithms) and their user interfaces (e.g., hands-free terminals). With the help of two independent auditory experiments with subsequent multidimensional analyses, three perceptual dimensions were identified for a diverse set of stimuli. These dimensions were labeled "directness/frequency content", "continuity", and "noisiness". Overall listening quality scores were collected in a separate experiment. A mapping of the obtained dimensions onto the overall listening quality scores by means of a linear model revealed that "continuity" appears to be the most important dimension in terms of overall listening quality.

Index Terms: assessment and modeling of speech quality, quality dimensions, multidimensional analyses

1. Introduction

Modern telephone connections show characteristics which differ from conventional ones with respect to the transmission techniques (packet-switched vs. circuit-switched), signal processing involved in the networks (e.g., simple codecs vs. noise reduction, echo cancellation, voice-activity detection, comfort noise injection), and the type of terminal equipment at the user's side (e.g., hands-free terminals or headsets vs. handset telephones). Thus, the characteristics of the transmitted speech signal may lead to new perceptual experiences by the listener compared to the perception of conventionally transmitted signals.

Quality is based on a subjective comparison of what is perceived with what is individually expected. This view on quality is reflected in the definition given by Jekosch [1]:

> "[Quality is the] result of judgment of the perceived composition of an entity with respect to its desired composition."

In this context, the considered entity is the transmitted speech perceived by the listener. During a reflective process, the listener decomposes the desired composition of the speech sound into desired features, and the perceived composition into perceived features [1][2]. Finally, the totalities of both the desired and the perceived features are internally compared. This leads to the overall listening quality judgment.

Perceptual quality dimensions related to traditional telephone channels, i.e. the features resulting from the decomposition, have been investigated quite thoroughly in the past [3][4][5]. Recently, the analysis has been extended towards the effects of mobile communication channels [6]. Still, several characteristics of modern channels have not yet been taken into account (e.g., packet loss, noise reduction). Because the physical characteristics of the modern connections are different from the ones of standard telephone connections, the perceptual effects will also be different. Therefore, different quality dimensions are expected, with different importance for the overall listening quality.

The awareness of these dimensions may lead to new approaches for instrumentally estimating the overall listening quality of a connection. The establishment of such an instrumental assessment is our long-term goal [7], since subjective experiments are expensive and time-consuming. So far, models are available which predict overall listening quality as perceived by the user, on the basis of measured speech signals [8] or of parameters describing the transmission channel [9]. Such models provide valid predictions for the channels they have been optimized for. However, they are not necessarily applicable to other channel characteristics, nor are they valid for predicting mouth-to-ear quality. It is expected that models which adequately consider the relevant quality dimensions will provide more generic predictions, i.e. predictions which are also valid for further developments of future technology. This paper is supposed to provide a basic reference in order to develop such a measure [7].

In order to identify the perceptual quality dimensions of modern telephone connections, two independent auditory experiments with subsequent multidimensional analyses were carried out, following different paradigms: multidimensional scaling (MDS) and semantic differential (SD). These experiments are presented in Section 3. A selection of stimuli representing all characteristics likely to be encountered in modern telephone networks was used in both experiments (Section 2). In a further experiment, the same stimuli were rated with respect to their overall listening quality (Section 4). The obtained quality scores can be predicted by combining the dimension scores of the respective stimuli.

2. Test stimuli and set-up

The speech files used in the experiments were degraded in a controlled way, using a circuit simulator for circuit-switched and packet-switched connections which is implemented on a programmable DSP system [2][10]. This system is able to generate degradations resulting, e.g., from codecs, band limitation, or circuit noise. Time-variant effects are generated by a scalable parametric packet-loss model, or by interrupting the channel with a cosine switch producing smooth ramps. In addition, the effects of two noise-reduction algorithms were considered, which may also show time-varying behavior (so-called "musical tones"). The user interface was modeled either as a typical handset-telephone filter, or realistic recordings were made with a head and torso simulator and a hands-free terminal (HFT). In the latter case, environmental noise with a low-pass frequency spectrum could be applied in the recording room.

A compromise has to be found with respect to the number of test stimuli, speakers and sentences. On the one hand all potentially relevant dimensions should be covered by the stimuli, so that the derived space should be valid for all imaginable speakers and spoken sentences. On the other hand, the effort for the test subjects increases enormously, especially in MDS. Here, $K·I·(I\text{-}1)$ judgments have to be made, where I is the number of stimuli and K is the number of speaker/sentence combinations. In order to make such experiments feasible anyway, K and I have to be relatively small, resulting in limited evidence for real-world scenarios (including the risk of underestimating between-speaker differences).

Two German sentences spoken by one male and one female speaker were selected as source material. Table 1 contains the fourteen conditions that were chosen as test scenarios. In order to obtain comparable results, the same set of stimuli was used in all experiments performed in this study.

Table 1: Connection characteristics

Abbrev.	Codec	Filter	Additional impairment
C1	G.711	handset	-
C2	G.726	handset	-
C3	G.729A	handset	-
C4	AMR	handset	-
H	G.711	flat	HFT
BP	G.711	handset	Bandpass 0.5-2 kHz
FL	G.711	flat	-
I1	G.729A	handset	10% packet loss
I2	G.729A	handset	20% packet loss
I3	G.711	handset	10% interruptions
HN	G.711	flat	HFT, room noise
HNR1	G.711	flat	HN, standard noise red.
HNR2	G.711	flat	HN, enhanced noise red.
CN	G.711	handset	Addit. circuit noise

The participants of the experiments were audiometrically screened. Each group represents a sample of telephone users, which, however, is not necessarily representative for the whole population, in particular with respect to the age (see descriptions below). All attendees were paid for their participation.

3. Multidimensional analyses

In order to reveal a mapping of the feature space of the listeners, two multidimensional analyses were carried out, following different paradigms with distinct advantages.

3.1 Multidimensional scaling (MDS)

The main idea of MDS is that distances between stimuli in the multidimensional feature space of individuals correspond to a mapped stimulus space [11]; here, this space is assumed to be Euclidean.

The distances can be obtained in a paired comparison test by collecting similarity judgments from the subjects. To this aim, a scale labeled with “very similar” and “not similar at all” at its extremities was used. By representing the similarity data in a multidimensional space, distinct points representing the stimuli can be found. The distances between them represent the degree of dissimilarity.

The dimensionality of the (reduced) space has to be selected so that the space adequately represents the observed stimulus distances. The goodness of fit of the space can be expressed by two parameters: The so-called S-stress (the lower the stress, the better), or the explained variance of the reduced space, R^2 (the higher the variance, the better) [11]. Furthermore, an important criterion for the choice of the dimensionality is the interpretability. Higher dimensionality increases the risk of modelling only the noise in the data.

The INDSCAL (Individual Differences SCALing) procedure used in this experiment additionally accounts for individual differences in judgments. Here, it is assumed that there are similar inter-individual feature spaces, i.e. spaces with equal dimensions but different individual weightings with regard to stimulus differentiation. By means of an adequate weighting of the individual spaces, a so-called group-stimulus space is calculated. For more details on the topic of MDS see [11], for example.

A group of 14 participants (6f, 8m), aged between 21 and 30, took part in this experiment. $2·I·\,(I\text{-}1) = 364$ pairs had to be judged by each subject. To avoid effects of fatigue, the experiment was divided into four sessions per subject. The order of stimulus presentation was randomized to avoid sequence effects.

3.2 Semantic differential (SD)

The advantage of MDS is the “unbiased” approach in the sense that no specific cues are given to the participants with respect to the features or characteristics of the speech samples. The drawback, however, is that the dimensions can only be interpreted by examining the configuration of the stimuli within the stimulus space. These drawbacks are avoided by the SD technique [12].

This experiment requires a pre-definition of a set of different descriptive attributes, or of pairs of opposite attributes (so-called antonyms). The test participants have to judge the occurrence and intensity of each attribute within a given stimulus on a bipolar scale, labeled with the respective antonyms. On the basis of the judgments for each stimulus, orthogonal factors can be derived with the help of factor analysis.

Due to the direct link between attributes and factors, the interpretation of the perceptual dimensions is largely facilitated. The main difficulty related to the SD in comparison to MDS is the a-priori determination of the attributes to be judged upon. In order to obtain an adequate set of attributes for the SD, pre-tests were carried out in order to find as many descriptive terms as possible for the given set of stimuli. This list was reduced in a second step to obtain a manageable number of judgments that should cover all effects that are perceptively important.

8 listeners (4f, 4m) who already took part in the MDS test, and thus had prior experience with the stimuli, were invited to spontaneously describe the stimuli by adjectives (e.g., “natural”), nouns (e.g., “naturalness”) or antonyms (e.g., “natural-unnatural”) or – if none of these types of words were found – another kind of description. The stimuli were judged in comparison with a reference, namely C1 (see Table 1).

A list of 217 different descriptions was collected. In a second pre-test, this list of descriptions was condensed to a limited set of attributes potentially suitable for the actual SD experiment. For this purpose, the most frequently named terms were carefully transformed into pairs of antonyms. Complementarily, relevant attributes regarding the stimuli were chosen by the experimenter based on findings in the related literature. Altogether, a list of 34 antonym-pairs was presented to the participants, who were asked to select the most relevant ones per stimulus. As a result, 13 antonym pairs were selected based on different criteria (e.g., overall frequency of selection, frequency of selection for a single stimulus): *interrupted-continuous*, *distant-close*, *crackling-not crackling*, *thin-thick*, *not noisy-noisy*, *muffled-not muffled*, *shaky-steady*, *indirect-direct*, *dark-bright*, *unintelligible-intelligible*, *not hissing-hissing*, *clear-unclear* and *distorted-undistorted* (translations from German wordings).

The use of a reduced set of 13 scales is also suitable with respect to effort and time. A number of $2 \cdot 13 \cdot I = 364$ judgments (both speakers considered) had to be made in two test sessions, without using a reference stimulus. To ensure that the meaning of the antonyms is the same for all participants, the antonyms were carefully described with corresponding synonyms.

18 participants took part in the final SD experiment (9f, 9m), aged between 21 and 31. They neither joined the pre-tests nor the MDS experiment.

3.3 Resulting quality dimensions

The (reduced) space derived from the SD was gained via a principal component analysis with Kaiser normalization and VARIMAX rotation. The SD shows no significant difference between the obtained spaces for both speakers (also indicated by ANOVA). Furthermore, Cronbach's alpha is 0.865, so the inter-subject reliability is very high. As a consequence, the collected values were averaged both over the speakers and the participants, resulting in a common multidimensional solution.

Three factors were extracted which can be interpreted both with the help of the correlated attributes (antonym pairs) and the configuration of the points (factor scores) representing the stimuli in this space (Figure 1). After rotation, the factors F_i cover a variance of 42.7% (F_1), 34.2% (F_2) and 16.6% (F_3).

F_1 seems to reflect the frequency content: At the negative end of F_1, both the bandlimited stimulus BP and the HFT-stimuli can be found, whereas all other stimuli are arranged more towards the positive end. Estimated transfer functions show that the HFT-related stimuli have a spectrum which strongly changes with frequency (comb filter effect), whereas the spectrum varies less over frequency for all other connections. The factor is highly correlated with *distant-close*, *indirect-direct* on the one hand, and *thin-thick*, *muffled-not muffled* and *dark-bright* on the other hand; "directness/frequency content" seems to be an adequate description.

F_2 was labeled with "continuity": At the positive end of F_2, "smooth" stimuli are accumulated. The negative end is made up by a cluster of the interrupted stimuli. Furthermore, the attributes *interrupted-continuous* and *shaky-steady* show high loadings on this factor. A certain degree of discontinuity can also be seen in the stimuli enhanced by noise reduction, in particular for the standard spectral subtraction algorithm (HNR1) which provokes "musical tones".

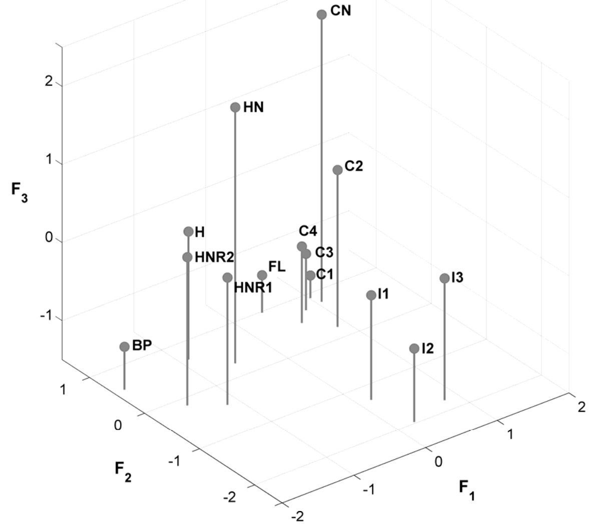

Figure 1: Stimulus space of the SD

Because of the displacement of the two noisy stimuli HN and CN and also C2 (that shows a slight signal-correlated noise), and high correlations with the antonyms *not noisy-noisy* and *not hissing-hissing*, F_3 is labeled "noisiness". Apparently, the noise reduction algorithms and the stimulus without additional noise and signal processing load equally high on this dimension. In particular, the "musical tones" of HNR1 are not perceived as noise.

It is striking that both the codec stimuli (C1-C4) and the interrupted stimuli (I1-I3) build clusters in all dimensions. Obviously, the respective stimuli are perceptively very similar.

The derived spaces for both speakers provided by MDS are similar to the one presented above, with two exceptions:

- A dimensionality of four was obtained by MDS (R^2=79.2%, S-stress=0.26 in the female case). For the MDS, F_1 of the SD is split into two separate dimensions "directness" and "frequency content". A three-dimensional solution (R^2=74.1%, S-stress=0.32) permits no reasonable interpretation.
- The spaces do not show such a high similarity between both speakers as in the SD. The interpretation of the results of both speakers is, however, the same.

Despite of the different dimensionality of the results, the interpretation of the derived spaces by SD and MDS remains the same. Due to the speaker-independency it was decided to consider the previously discussed three-dimensional solution of the SD in the remainder of this paper.

4. Modeling of overall listening quality

Overall listening quality judgments of the same set of stimuli were collected in a separate listening test to investigate the importance of the dimensions regarding the overall listening quality. The subject group of MDS was employed for this test, which, however, took place prior to the MDS experiment. Thus, the subjects were still untrained at this point in time.

In order to generalize the judgments across the speakers, two additional speakers were considered. Thus, $4 \cdot I = 56$

conditions had to be judged per subject. The stimuli were presented in randomized order.

To avoid known drawbacks of the classical procedure of absolute category rating (low resolution and saturation effects), a continuous scale was used which is described in [10]. The computed arithmetic means (*mean opinion scores*) over all subjects are abbreviated by MOS_C. The used scale was labeled with (corresponding numerical values in brackets) "extremely bad" (<1), "bad" (1), "poor" (2), "fair" (3), "good" (4), "excellent" (5) and "ideal" (>5).

The calculation is performed via a multivariate linear regression. The MOS_C values across subjects and across those speakers that were also judged in the experiments described in section 3 are chosen as the target variable. The coordinates of the points representing the stimuli in the space are the predicting variables. The model follows the relation

$$MOS_C = const. + \sum_{i=1}^{3} b_i F_i \tag{1}$$

and covers about 90% of the total quality variance.

The Z-score normalized regression coefficients b_i determined in this way are b_1=0.46 ("directness/ frequency content"), b_2=0.70 ("continuity") and b_3=–0.47 ("noisiness"). These values can be interpreted as correlation coefficients of the regression line with the respective dimension.

Because b_2 shows the highest absolute value amongst the coefficients, "continuity" is the most important feature for modeling the overall listening quality (the more continuous a stimulus is, the better the quality). Since an ANOVA revealed that besides the factor *stimulus* also *speaker* is significant, the calculation of MOS_C values across speakers may not be justified. But except for a shift of weightings for the other factors, "continuity" remains the most important dimension in terms of overall listening quality. This also holds true when using the MDS solution as predicting variables. Therefore, "continuity" can be considered as a speaker-independent relevant dimension for overall listening quality.

5. Conclusions and outlook

Two auditory experiments with subsequent multidimensional analyses following different paradigms have been carried out in order to extract the underlying quality dimensions of modern telephone connections. Three speaker-independent dimensions could be identified. These dimensions were labeled "directness/frequency content", "continuity", and "noisiness". Separately collected overall quality scores could be predicted assuming a linear model of these dimensions. In accordance with [6] which was carried out in a mobile context, "continuity" seems to be the most important dimension when channels of time-varying characteristics are involved. Because interruptions primarily have an impact on speech intelligibility and listening-effort, both seem to be dominant aspects of overall speech quality (cf. e.g. [2]).

Another important result is that the used codecs were found to be very similar, with respect to all three dimensions. Thus, multidimensional investigations which have been performed in the past [3][4] represent only a small region of the total perceptual feature space. In our case, these perceptually similar stimuli could not be differentiated and consequently play only a subordinate role for describing the perceptual space.

All in all, it can be said that a large range of possible effects has to be considered to assess quality of transmitted speech. This must be done for the whole chain mouth-to-ear, considering all impacting elements of the transmission channel. Although the set of stimuli and speakers used in this study were limited due to the experimental procedure, the dimensions constitute a basis for more detailed analyses. In order to increase the resolution of single dimensions and validate them for different speaker/sentence combinations, further stimuli are currently generated and assessed with regard to both their perceptual attributes and overall listening quality. This will enable the optimization of the diagnostic listening quality prediction model we aim at [7]. Ultimately, it is our goal to quantify the perceptual dimensions based on instrumentally measurable physical correlates.

6. Acknowledgements

The present study was carried out at IKA, Ruhr-Universität Bochum (Prof. R. Martin, Prof. em. J. Blauert, PD S. Möller). It was supported by the Deutsche Forschungsgemeinschaft (DFG), grant MO 1038 and HE 4465.

7. References

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

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

[3] Bappert, V., Blauert, J., "Auditory Quality Evaluation of Speech-Coding Systems", in: *acta acustica*, 2, pp. 49-58, 1994.

[4] Hall, J., "Application of Multidimensional Scaling to Subjective Evaluation of Coded Speech", in: *J. Acoust. Soc. Am.*, 110(4), pp. 2167-2182, 2001.

[5] McDermott, B., "Multidimensional Analyses of Circuit Quality Judgments", in: *J. Acoust. Soc. Am.*, 45(3), pp. 774-781, 1969.

[6] Mattila, V.-V., *Perceptual Analysis of Speech Quality in Mobile Communications*, PhD thesis, Tampere University of Technology, Publ. 340, 2001.

[7] 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. 4th European Congress on Acoustics (Forum Acusticum 2005)*, H-Budapest, 2005.

[8] ITU-T Rec. P.862, *Perceptual Evaluation of Speech Quality (PESQ),* International Telecommunication Union, CH-Geneva, 2001.

[9] ITU-T Reg. G.107, *The E-model,* International Telecommunication Union, CH-Geneva, 2005.

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

[11] Kruskal, J., Wish, M., "Multidimensional Scaling", Vol. 07-011 of *Quantitative Applications in the Social Sciences (E.M. Uslaner, ed.)*, Sage, Newbury Park, 1978.

[12] Osgood, C.E., Suci, G., Tannenbaum, P., *The Measurement of Meaning,* University of Illinois Press, Urbana, 1957.