Perceptual Dimensions of Wideband-transmitted Speech

Marcel Wältermann[1], *Alexander Raake*[2], *and Sebastian Möller*[2]

[1]Institute of Communication Acoustics, Ruhr-University Bochum, Germany
[2]Deutsche Telekom Laboratories, TU Berlin, Germany
marcel.waeltermann@ruhr-uni-bochum.de

Abstract

In this paper it is analyzed which perceptual dimensions are existent for speech that is transmitted over wideband telephone connections. Therefore, two auditory experiments with subsequent multidimensional analyses (multidimensional scaling and semantic differential) were carried out with a diverse set of mixed narrowband and wideband conditions. This revealed a mapping of the perceptual space with a dimensionality of four. The identified dimensions were labeled "continuity", "distance", "lisping", and "noisiness". By mapping those dimensions onto overall listening quality scores in a linear way, statements regarding the importance of single dimensions could be made for the given context. It turned out that "continuity" is the main contributor for overall listening quality.

1. Introduction

The trend towards speech transmission over packet-based networks leads to a new level of quality experienced by the user [1]. The universal technique of the underlying packetization of any media content imaginable, and thus also speech, permits the disengagement of physical restrictions given by traditional public-switched networks, or mobile networks of the second generation. In particular, the constraint of audio-band limitation can be abandoned. Hence, the former limitation of 300-3400 Hz is replaced by a wider band, e.g. 50-7000 Hz.

Although the speech quality is improved by wideband transmission at first glance, there are several elements of the transmission chain mouth-to-ear that potentionally degrade the overall quality. The impact of these elements on narrowband speech quality has been investigated thoroughly in the past (cf., e.g., [1][2]). But since both the wideband transmission itself and degradations of wideband-transmitted speech may lead to an unconventional quality experience for the user, aspects influencing the speech quality have to be reconsidered for the wideband case.

Ambient noise, e.g., which is transmitted through the wider band, may be perceived as more annoying in comparison to narrowband noise. The auditive effects of signal processing units such as noise suppressors, echo cancelers or packet loss concealment (PLC) algorithms may be perceived differently in the wideband case. If packet-based networks are interconnected with narrowband public-switched networks, effects like circuit noise, codec tandeming, or additional echo effects emerge. In this common scenario, wideband transmission is even entirely impossible. However, the signal at the receiver's side may artificially be regenerated by bandwidth enhancement techniques. Furthermore, considering the user device, most terminals inadequately transmit a wider frequency range due to physical restrictions.

In order to ensure a certain quality of the voice-service, overall quality predictions are adjuvant for network providers in both the planning phase and in-service of a network. For this reason, parameter-based [3] and signal-based [4] measures for narrowband speech have been established. Such instrumental measures aim at predicting the judgments that would have been made by participants in auditory tests. The measures provide satisfying results for channels they have been designed for. Recently, efforts have been made to adapt these measures to wideband speech [5]. However, the latter suffer from certain inconsistencies to results gained from auditory experiments [6].

A more general approach to quality prediction constitutes a perceptually motivated measure, based on quality-relevant dimensions [7]. These dimensions reflect the perceptually relevant characteristics of the speech signal that a user takes into account when judging the overall quality. The knowledge of wideband dimensions in telephony is particularly interesting since up to now, no investigations known to the authors have been made in this direction. The dimensions allow to diagnose the overall quality judgments, i.e., to reveal the reasons of the judgments. Two speech samples, e.g., that yield the same overall quality judgment may show different perceptual characteristics. By estimating the dimensions on the basis of the physical signal and by applying an appropriate mapping function, overall quality can be predicted. Since the mapping function can be adjusted in order to incorporate new effects, the approach is promising to be valid also for future technologies, assuming that the dimensions remain the same.

The present paper addresses the fundamentals necessary in order to develop an instrumental model for wideband-transmitted speech based on perceptual dimensions, namely the identification of the respective dimensions. This is done by two independent auditory experiments with subsequent multidimensional analyses, which then provide an approximation of the actual perceptual space. The two experiments follow different paradigms: Multidimensional scaling (MDS) and semantic differential (SD). Both methods as well as the analyses of the experiments will be described in Section 3.

In order to ensure that all perceptually relevant dimensions are captured, a comprehensive set of speech samples is required containing different effects likely to be encountered in mixed narrowband/wideband networks: Different narrowband and wideband codecs, bandpasses, circuit and ambient noise, noise suppression, packet loss with PLC, and artificial bandwidth enhancement. Technical details about the processing of the speech samples are given in Section 2, together with a description of the test set-up common in both experiments.

In Section 4, it is assumed that overall listening quality can be modeled in a linear way by means of the results of the derived perceptual space. Therewith, the "importance" of distinct dimensions in terms of overall quality can be quantified.

Conclusions are drawn and an outlook towards further work is given in Section 5.

2. Speech samples and experimental set-up

Two German sentences (one male and one female speaker), taken from [8], were chosen as the source material. These are the same sentences already used in [9], so that the perceptual space for narrowband speech therein described can directly be compared to the space derived in this paper.

The speech files were mainly processed by means of a circuit simulation tool for both circuit-switched and packet-switched networks (all noise sources were switched off in order to avoid a predominant noise effect in the samples which potentially could mask the other effects). By means of this simulation tool, which was developed in [2] and extended in [1], a diverse set of wideband (WB) speech samples could be produced. The basis is constituted by "clean" samples which are only limited regarding their audio bandwidth (50-7000 Hz). These samples represent certainly the highest quality level possible within the whole set. Further samples are generated including different wideband codecs (e.g., Adaptive Multi-Rate Wideband, AMR-WB) and a bandpass filter. To simulate a hands-free situation on sender's side, samples are re-recorded by means of a talking head-and-torso simulator and an ideal hands-free terminal (HFT), introducing an additional acoustic path. Background noise (so-called Hoth noise) is introduced in the same scenario, and also a noise suppression procedure. Due to the simplicity of the algorithm (spectral subtraction), so-called "musical tones" are clearly noticeable in these samples. Since the simulation tool is yet not capable to process wideband data over the packet-switched part, the error insertion device of [10] is used in order to introduce random frame erasures (assuming that an IP packet contains only a single frame, frame loss rate then equals the packet loss (PL) rate). Each packet contained a 20 ms frame of the speech signal.

Apart from wideband speech samples, also narrowband (NB) samples (300-3400 Hz), taken from [9], with diverse kinds of degradations are included. This takes account of the fact that for the time-being, both narrowband and wideband speech is most likely to be encountered in real life. In addition, the bandwidth of one of the narrowband sample was artificially extended (artificial bandwidth enhancement, ABE). This could be achieved by applying an algorithm that extrapolates the frequency envelope towards higher frequencies by a LPC-codebook approach [11].

Since the effort of the experiments increases rapidly with the number of samples, $I = 14$ processing chains were chosen as a compromise between variety of stimuli and effort. These conditions are listed in Table 1.

The room in which the experiments were performed does not exactly meet the requirements given in [12], but was considered as sufficient with regard to the noise level. For the tests, a diffuse-field equalized STAX Lambda Pro headset was used. The stimuli were presented diotically, the listening level was fixed to 73 dB SPL. The participants were students from the local university (aged between 18 and 34); each of them got paid.

Table 1: *Processing chains*

Abbreviation	WB/NB	Processing elements
CLEAN	WB	direct channel
G7221	WB	G.722.1@24kbps
AMRWB	WB	AMR-WB@6.6kbps
G711	NB	G.711
BP_N	NB	G.711, 0.5-2 kHz bandpass
BP_B	WB	0.1-5 kHz bandpass
HFT_NB	NB	HFT
HFT_WB	WB	HFT
NC	NB	G.711, circuit noise
HFT_WB_N	WB	HFT, background noise
HFT_WB_NR	WB	HFT, noise suppression
PL20_NB	NB	G.729A, 20%PL
PL20_WB	WB	AMR-WB@23.05kbps, 20%PL
ABE	WB	G.711 codec, art. bandw. enh.

3. Multidimensional analyses

In order to reveal a mapping of the perceptual space of the listeners, two multidimensional analyses were carried out, following different paradigms with distinct advantages. The methods and experiments are described in this section, as well as statistical analyses of the experimental data and the resulting perceptual dimensions.

3.1. Multidimensional scaling

A multidimensional scaling experiment (MDS) [13] was performed with 19 participants (9 f, 10 m). They were asked to judge the (dis-)similarity of pairwise presented speech samples. To this end, a scale labelled with "very similar" and "not similar at all" at its extremities was used. The main idea of MDS is the mapping of the dissimilarity to a metric distance, i.e., a large distance between two points representing the speech samples corresponds to a high dissimilarity. This ultimately results in a n-dimensional configuration of points. Being I the number of processing chains, a total of $2I(I-1)$ pairs (two speakers) had to be judged in four sessions by each participant. Before each session, participants could get familiarized with the task judging some training samples.

To take into account individual differences in judgments, INDSCAL (INdividual Differences SCALing) was selected as the calculation method. 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 and the INDSCAL approach, see [13], for example.

The dimensionality n is determined considering both statistical goodness-of-fit parameters and the ability to interpret the dimensions. Prominent parameters are the Stress (goodness-of-fit between judgments and distances; the lower the stress, the better) and the covered variance of the model, R^2 (the higher the covered variance, the better).

3.2. Semantic differential

The strength of MDS is that no specific cues about features of the stimuli are given to test participants. Plus, judging the similarity is a relatively simple task. However, due to the lack of additional hints, the determination of the dimensionality and the interpretation of the results is quite difficult for the experimenter and can eventually be done only intuitively. Therefore, a semantic differential experiment (SD) [14] was additionally conducted in order to facilitate the interpretation.

The paradigm of SD is rather contrary to MDS since in SD,

a predefined set of dimensions is given to the test participants in terms of bipolar scales. The extemities of each scale are labeled with pairs of opposite attributes, so-called antonyms, each describing a one-dimensional feature. The occurrence and intensity of each feature within a given stimulus have to be judged by test participants in an absolute way (i.e., without a reference). Each stimulus can thus be arranged in a m-dimensional space, where m is the number of scales. With help of factor analysis, the dimensionality m can be reduced in order to reveal the $n < m$ latent perceptual dimensions, in form of orthogonal factors.

The main difficulty in SD, however, is the a-priori determination of the descriptive attributes. They should also capture possibly the whole perceptual space, in a redundant manner. In order to find adequate attributes, two extensive pre-tests were conducted. In pre-test 1, as many descriptions as possible were collected by 10 participants (5 f, 5 m) who already took part in the MDS experiment and therefore were familiar with the stimuli. The terms were supposed to be 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 participants were also asked to give a rating of the intensity of each term to ensure that participants include terms for subtle features as well.

This procedure resulted in 135 different descriptions (altogether, 1073 terms were collected). The desciptions were weighed by frequency of occurrence in both overall and in single stimuli. The most frequently named terms were carefully translated into antonym-pairs and presented to the participants in the second pre-test.

The purpose of this test was to find a common agreement by the participant group upon the antonyms individually found and translated from pre-test 1. The task was to select those antonyms for each stimulus that were regarded as perceptively significant. Again, a weighting rule was applied and as a result, a set of 28 antonym-pairs were finally selected for the actual SD experiment:

Direct–indirect, *hissing–not hissing*, *thin–thick*, *regular–irregular*, *nasal–not nasal*, *rough–smooth*, *close–distant*, *reverberant–dry*, *brassy–not brassy*, *clear–unclear*, *spatial–not spatial*, *muffled–not muffled*, *chopped–not chopped*, *blurred–not blurred*, *doubled–not doubled*, *bubbling–not bubbling*, *crackling–not crackling*, *tight–wide*, *noisy–not noisy*, *husky–not husky*, *metallic–not metallic*, *lisping–not lisping*, *clinking–not clinking*, *crispant–fuzzy*, *warm–cold*, *ragged–not ragged*, *creaking–not creaking*, and *rattling–not rattling* (translations from German wordings).

Two different groups of participants took part in the SD experiment. Group 1, consisting of 9 students (5 f, 4 m), had also joined the MDS experiment and the pre-tests. They can therefore be considered as experts in the sense that they were aware of both the stimuli and commonly agreed upon the antonym-pairs. Group 2, consisting of 19 students (8 f, 11 m), had not taken part in any of the previous experiments, and were therefore naïve. Nevertheless, both groups took part in a preceding training session, consisting of listening to all stimuli in a row, and performing test judgments. In order to make the meanings of the antonyms more clear, each of them was described by synonyms.

Considering both speakers, a number of $2 \cdot 28I$ judgments in two separate sessions have thus to be made per participant. Two subjects did not appear to the respective second session, so that judgments for one male speaker and one female speaker were missing.

3.3. Pre-analyses of the MDS and SD data

There are some decisions to be made in order to subsume data obtained from the SD experiment before the factor analysis can be performed. An ANOVA for repeated measures with the three within-subject factors *scale*, *speaker* and *processing chain* and the between-subject factor *subject-group* revealed that the latter is not statistically significant ($F = 1.47, p = 0.24$). Hence, both groups are analyzed conjointly in the following. Further, the within-subject factor *speaker* is not significant either ($F = 0.46, p = 0.51$). This justifies the calculation of the mean values over both speakers. This is also supported by very high correlations between extracted factors of different solutions (i.e., single speaker vs. mean over speaker).

The data now consist of three modes, representable by a three-way array *subject*×*processing chain*×*scale*. Ordinary factor analysis, however, allows only two modes to be considered (namely cases and variables). Therefore, the authors decided to calculate the arithmetic mean over the *subject* mode, resulting in a *processing chain*×*scale* matrix. This is a popular approach when analyzing SD data. However, this is not necessarily justified, since there might be individual differences that are disregarded in doing so. These differences would perhaps lead to a different, more fine-grained solution. Additionally, they would give evidence how generalizable the results are. Even though there are some examples of three-mode factor analysis of SD data in the literature, this approach is neither widely spread nor trivial in both application and interpretation. A re-analysis of the available data in a later publication, however, is reserved. In the present paper, the two-mode SD data is reduced in dimensionality by principal component analysis (PCA) and VARIMAX rotation.

Although an ANOVA for repeated measures reveals that the within-subject factor *speaker* is not relevant for the MDS data ($F = 1.48, p = 0.24$), the resulting dimensions differ more severely as compared to the factors of the SD data when using the mean over both speakers. The reason might be that the INDSCAL procedure is quite sensitive even with respect to small changes in the similarity data. Nevertheless, the interpretation of both solutions (male and female speaker) is the same. In the following, the MDS solution for the female speaker will be analyzed exemplary.

In order to determine an appropriate dimensionality n, statistical parameters of both the MDS and SD spaces are investigated. Starting with MDS, it is important to find an adequate trade-off between goodness-of-fit and interpretability of the resulting space. It turned out, regardless of the results of the SD, that a 4-dimensional space is well interpretable, whereas an explanation of a 5-dimensional solution is hardly possible. Kruskal's Stress–1 [13] of a 4-dimensional solution is $Stress = 0.19$, the covered variance in this case is $R^2 = 75\%$. So in terms of the goodness-of-fit measures, there would still be some headroom left to increase the dimensionality.

The dimensionality of $n = 4$ is also supported by investigating the eigenvalues of the correlation matrix calculated by PCA of the SD data. Taking a fifth factor would result in a corresponding eigenvalue < 1, meaning that this factor would cover less variance of the data than a single scale (so-called Kaiser criterion). The factors F1 to F4 of the VARIMAX-rotated 4-dimensional space cover a variance of 33.1%, 21.8%, 21.2%, and 17.2%, respectively.

After reducing the dimensionality by means of factor analysis, the scales (i.e., the attributes) are more or less highly correlated with the resulting factors (these correlations are called

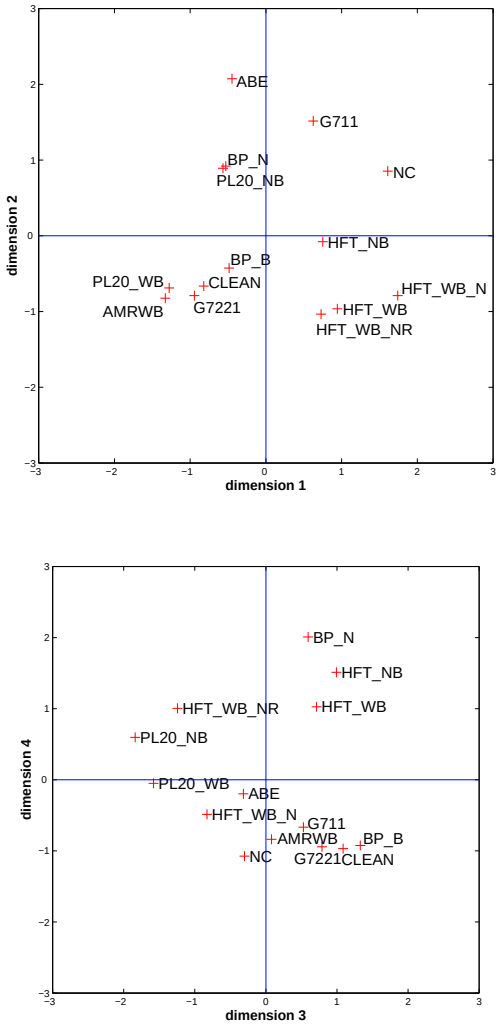

Figure 1: *Mapping of the stimuli in the perceptual space, obtained by MDS (female speaker).*

factor loadings). Factor loadings that are higher than a value of 0.7 are discussed in the analysis of the next section. The value of each stimulus on a common factor is called *factor score*. As in MDS, the factor scores, representing the speech samples, can be arranged in the space spanned by the orthogonal common factors. In order to compare the point-configurations of the speech samples between the MDS and SD solution, Pearson's correlations between single axes can be calculated. This will be done and discussed in the next section. (Note that the numbering of the factors do not necessarily correspond to the numbering of the dimensions in terms of interpretation.)

3.4. Resulting quality dimensions

In Figure 1, the configuration of points representing the stimuli in the 4-dimensional space derived by MDS is depicted for the female speaker in a simplified way.

The positive end of dimension 1 is made up by the noisiest stimuli in the set, namely HFT_WB_N and NC. With decreasing values of this dimension, the noise in the stimuli increases. Compared with the SD solution, this dimension is highly and significantly correlated ($r = 0.83, p < 0.01$) with F3, which in turn shows high loadings of the attributes *noisy*, *blurred*, and *hissing*. This dimension can therefore clearly be labeled with "noisiness". Obviously, the HFT condition without background noise in the wideband case (HFT_WB) is perceived as slightly more noisy than in the narrowband case. However, the relatively poor noise suppression algorithm enhances the respective speech sample sufficiently. Further, also the signal-correlated noise inherent in G.711-coding plays a role in this dimension.

The most positive point in dimension 2 is the one representing the artificially bandwidth-enhanced stimulus, followed by stimulus G711 and other narrowband stimuli of various distortions. All wideband stimuli are arranged more densely at the negative end. The attribute *lisping* shows the highest loading on the correlated factor, followed by *clinking* and *rattling* (correlation between factor F4 and this dimension $r = 0.71, p < 0.01$). Roughly speaking, lisping is the disability to pronounce sibilants properly, and instead replace them with interdentals. In fact, the effect of lisping is most conspicuous for stimulus ABE. Since sibilants usually exhibit considerable energy in higher frequency regions, as compared to other speech sounds, this effect apparently originates from the imperfect enhancement of the higher frequency band. The narrowband stimuli which lack higher frequencies do not render sibilants properly as well, especially compared to wideband stimuli. Following this notion, their location with regard to this dimension can then be explained well. The label "lisping" can thus be interpreted as an anomaly or lack of high frequency components which are necessary for rendering sibilants correctly.

One additional effect is striking with regard to this dimension: Since all wideband stimuli are arranged on the negative end all narrowband stimuli (besides ABE) are arranged on the positive end, this dimension roughly classifies "natural" wideband and narrowband stimuli. The artificial wideband stimulus ABE belongs to the narrowband stimuli in this respect.

Stimuli showing time-varying behavior can clearly be differentiated in the third dimension (negative end). In particular, both stimuli which are distorted by packet loss exhibit the strongest effect in this sense. But also the "musical tones" produced by the simple noise suppression algorithm as applied to HFT_WB_N has a considerable effect in a time-varying manner. There is a high correlation of this dimension with F1 ($r = 0.75, p < 0.01$), which in turn is loaded by *bubbling*, *irregular*, *chopped*, *ragged*, *fuzzy*, and *unclear*. Apparently, this dimension reflects the "continuity" of the signal.

Except for HFT_WB_N, all stimuli processed with HFT can be found at the positive end of dimension 4. These stimuli contain the influence of the additional acoustical path from the artificial mouth of the head-and-torso simulator to the microphone of the HFT. They therefore sound more indirect. This is reflected by a high loading of the attribute *indirect* on the corresponding factor F2 (correlation between dimension 4 and F2 $r = 0.69, p < 0.01$). Surprisingly, the attribute *spatial* is negatively correlated, meaning that participants judged the respective stimuli as *not spatial*. This was not expected and is somewhat puzzling, since it was believed that, e.g., the reflections of the room have an impact on this dimension (cf., [9]). Nevertheless, the high loadings of *tight*, *nasal*, *distant*, *thin*, along with the dominance of the heavily bandlimited stimulus BP_N, indicate that certainly the tightness, or the distance are of major importance. Perhaps, *not spatial* was interpreted by "far away". This dimension can then be labeled with the term "distance".

In the narrowband-only case [9], a comparable dimension

was labeled with "directness/frequency content". In the mixed narrowband/wideband case presented here, a labeling with "frequency content" is not justified anymore, since both the positive and negative part of this dimensions can potentially occupied by both narrowband and wideband stimuli. A narrowband/wideband classification is rather reflected by dimension 2.

Finally, Figure 1 shows that with respect to all dimensions, the stimuli CLEAN, G7221, AMRWB, and BP_B constitute a cluster. Apparently, neither those codecs nor the relatively weak bandwidth limitation of BP_B are distinguishable within the perceptual space.

As mentioned before, the solution of the male speaker is similar, though not identical to the one of the female speaker. However, the interpretation, i.e., the labeling of the dimensions, is the same.

4. Modeling of overall listening quality

In a further experiment, the overall quality of each speech sample was rated by those participants who already took part in the MDS experiment. Since this test took place prior to the MDS experiment, all participants were still unbiased at this point in time. In order to meet requirements given by [12], two additional speakers were considered in this test. Via multivariate linear regression, the dimension scores derived by MDS were mapped onto the mean overall quality scores (mean opinion scores, MOS), following the equation

$$\text{MOS} = \text{const.} + \sum_{i=1}^{n} b_i \cdot dim_i \,, \qquad (1)$$

where dim_i are the coordinates of a speech sample in the space, b_i are weighting factors, and n is the dimensionality with $n = 4$. Since an ANOVA of the MOS data exhibits a strong speaker dependency, the MDS solution of the female speaker is mapped onto the MOS scores of the same speaker here.

The model covers a variance of about $R^2 = 75\%$. The magnitudes of the coefficients b_i indicate the importance of the respective dimension with regard to overall quality, showing that the dimension "continuity" is apparently of major importance in this context ($b_3 = 0.78$). This is also somehow reflected by the high value of the covered variance of F1 (see Subsection 3.3). "Noisiness" exhibits a weighting coefficient of $b_1 = -0.13$ and seems to be of minor importance. Interestingly, neither of both frequency-related dimensions "distance" ($b_4 = -0.30$) and "lisping" ($b_2 = -0.14$) are considerably important. The introduction of wideband speech alone is obviously not necessarily enough in order to provide satisfying mouth-to-ear speech quality. It is more important to consider the additional degradations which this may have as a consequence.

The weighting coefficients change if, e.g., the solution of the factor scores of the SD solution are taken as the predicting variables. However, continuity remains the most important dimension.

5. Conclusions and outlook

Two different auditory experiments with subsequent multidimensional analyses were carried out in order to reveal a mapping of the perceptual space in the context of wideband-transmitted speech (50-7000 Hz). Both experiments resulted in mappings that highly resemble each other. By investigating several goodness-of-fit measures and comparing the results of both experiments, it turned out that a dimensionality of $n = 4$ seems to be stable in terms of interpretability. By modeling overall listening quality from those dimensions, statements regarding the importance of each dimension in the given context could be made. The following dimensions could be identified (in the order of their importance for overall listening quality):

- Continuity,
- distance,
- lisping, and
- noisiness.

The results are comparable to findings given in [9], where a three-dimensional space was derived for narrowband-only speech (300-3400 Hz). Here, however, the space is extendend by a further dimension which can be labeled with "lisping". It is assumed that this dimension reflects effects that are physically located in the higher frequency range. This could be the reason why this dimension is hidden or not existent at all in the pure narrowband case.

Although differences in the individual judgments were taken into account in the MDS analysis, this was not the case for the factor analysis of the SD data, due to practical reasons. This may have ruled out some subtle information, and perhaps even a further dimension. Goodness-of-fit parameters would also justify a higher dimensionality for the MDS, so that a further dimension could be latent in the similarity data. This, in turn, should be supported by the SD in order to interpret the space properly. Therefore, so-called three-way factor re-analysis of the SD data should be carried out in the future. This would probably not change the strong main dimensions, but may reveal an additional, more subtle dimension.

The fact that "continuity" is a main contributor for overall quality in the wideband case gives evidence that the transition from narrowband to wideband is not necessarily enough in order to provide a better quality for telephony. Rather, the whole transmission chain from the mouth of the speaker to the ear of the listener should be regarded to take into account all potentially quality-affecting elements. Here, the different codecs or weak band-limitations were found to be of minor impact, since they represent only a small region in the perceptual space.

In order to develop a valid model for predicting overall quality (which does not necessarily have to be linear), the relation between the dimensions and overall quality have to be analyzed more thoroughly. In a first step, this could be done by increasing the resolution of single dimensions by investigating them separately. For this reason, also different speaker/sentence combinations should be taken into account.

The ultimate aim is the development of an instrumental measure that is based on perceptual dimensions. In the process of finding physical correlates of the dimensions, the obtained results should provide an adequate means. It is assumed that this process is facilitated by the fact that approaches for measuring similar dimensions for narrowband speech are currently being developed [15], and can be extended towards the measurement of wideband dimensions.

6. Acknowledgement

The present study was carried out at Deutsche Telekom Laboratories, TU Berlin, Germany. It was supported by the Deutsche Forschungsgemeinschaft (DFG), grant MO 1038. The authors would like to thank Julia Fedorova and Klaus-Peter Engelbrecht who carried out the experiments.

7. References

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

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

[3] ITU-T Rec. G.107, *The E-Model, a Computational Model for Use in Transmission Planning*, International Telecommunication Union, CH-Geneva, 2005.

[4] 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 an Speech Codecs*, International Telecommunication Union, CH-Geneva, 2001.

[5] ITU-T Rec. P.862.2, *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, International Telecommunication Union, CH-Geneva, 2005.

[6] Takahashi, A., Kurashima, A., Morioka, C., Yoshino, H., "Objective Quality Assessment of Wideband Speech by an Extension of ITU-T Recommendation P.862", in: *Proc. 9th European Conference on Speech Communication and Technology (Interspeech 2005)*, ES-Lisboa, pp. 3153-3156.

[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)*, HU-Budapest, pp. 1695-1700.

[8] ITU-T Rec. P.501 Amendment I, *Test Signals for Use in Telephonometry*, International Telecommunication Union, CH-Genf, 2004.

[9] Wältermann, M., Scholz, K., Raake, A., Heute, U., Möller, S., "Underlying Quality Dimensions of Modern Telephone Connections", accepted for *International Conference on Spoken Language Processing 2006*, US-Pittsburgh, PA, 2006.

[10] ITU-T Rec. G.191, *Software Tools for Speech and Audio Coding Standardization*, International Telecommunication Union, CH-Geneva, 2005.

[11] Carl, H., Heute, U., "Bandwidth Enhancement of Narrow-Band Speech Signals", in: *Proc. VII. European Signal Processing Conference (EUSIPCO 1994)*, UK-Edinburgh, Scotland, pp. 1178-1181.

[12] ITU-T Rec. P.800, *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, CH-Genf, 1996.

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

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

[15] Scholz, K., Wältermann, M., Huo, L., Raake, A., Möller, S., Heute, U., "Estimation of the Quality Dimension 'Directness/Frequency Content' for the Instrumental Assessment of Speech Quality", accepted for *International Conference on Spoken Language Processing 2006*, US-Pittsburgh, PA, 2006.