euronoise

How can we model a quality event? Some considerations on describing quality judgment and prediction processes

S. Möller

Deutsche Telekom Laboratories, Berlin Institute of Technology, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
sebastian.moeller@telekom.de

Blauert has introduced a systemic view on a listener in an auditory experiment. This view helps to separate sound events from auditory events and from their descriptions, and to identify and describe the processes involved in such experiments. This notion has been extended to listeners in a quality-judgment situation by Jekosch and Raake, leading to the notion of a "quality event". This paper will identify components which are necessary for an algorithmic description of the processes involved in the formation of a quality event. Taking the example of telecommunication services, it will be shown which components of quality prediction models are already available, and which others are still out-of-reach and require further study.

1 Introduction

With the increasing development of communication technology, the need for evaluating their quality becomes urgent. The discipline of communication acoustics is not an exception to this rule: Acoustics engineers need to quantify the quality of, e.g., telephones, public announcement systems, concert halls, or car noises, in order to design optimum systems for the end user. Having said this, it comes to a surprise that the definition of quality – in particular in the context of sound quality – did not prove to be stable; in fact, it has radically changed during the last 10 years.

At about that time, quality was considered to be the "totality of characteristics of an entity […] that bear on its ability to satisfy stated or implied needs" (EN ISO 9000, 2000, cited after [10]). This "totality of characteristics" is what we would nowadays call the 'character' of an entity [2]. The quality definition has been improved in 2005, now stating that quality is the "degree to which a set of inherent characteristics […] fulfils requirements" [7]. It is further stated that a characteristic is a "distinguishing feature", and the term 'inherent' is used explicitly as opposed to 'assigned', meaning "existing in something, especially as a permanent characteristic" [7]. Although being improved, this definition is not yet in line with the framework developed by Blauert and Jekosch, in particular when it comes to sound quality.

According to their framework of definitions, quality is defined from a perceiving person's point-of-view. Jekosch [10] defined quality as follows:

> "Result of judgment of the perceived composition of an entity with respect to its desired composition."

Apparently, quality involves a perception and a judgment process, during which the perceiving person compares the perceptual event with a (so-far unknown) reference. The character of the perceived composition is not necessarily a "permanent characteristic" of an entity; in fact, the reference may influence what is actually perceived. In any case, as the result of the comparison, quality is always *relative* and happens as a 'quality event' in a particular spatial, temporal and functional context. Such a context has to be modelled when quality is to be quantified through a measurement process.

In order to analyze quality and to design high-quality systems, knowledge about the perception and judgment processes involved in the formation of a quality event is necessary. The following section will review some of the processes which have been postulated so far, starting from a systemic approach to a listener in an auditory experiment introduced by Blauert in 1974 [5], and extending it to an interactive person in a quality judgment experiment.

The processes may serve as 'building blocks' in two ways: First, knowledge of the involved processes is necessary to design appropriate measurement processes for, e.g., sound quality, transmission quality, auditory-scene quality, or product-sound quality. Secondly, the blocks enable us to define algorithms which estimate quality – or sub-aspects of it – in the system design phase. Section 3 will follow the second path and try to define the components of a future general model which might be able to predict the quality assigned by an interactive person. Section 4 will review which parts of such a model are already available for telecommunication services, and which others are still out-of-reach. Finally, a list of open questions and some personal opinions are given in Section 5.

2 Quality judgment and prediction

As stated above, quality formation requires a perceiving and judging person to be available. Although such perception and judgment processes happen implicitly in everyday-situations, it is desirable that these processes can be provoked in a well-defined and mostly-controlled context when quality is to be quantified. This usually happens in a perceptual experiment. For the acoustic modality, Blauert [4] introduced a systemic representation which sketches the involved processes. This representation is reproduced in Fig. 1.

Input to the "system" (i.e. the listener) is a sound event s_0 reaching the listeners ear(s). The sound event may give rise to a perceptual event (here: auditory event h_0) happening inside the listener. Unfortunately, this event is not accessible to the experimenter: He has to ask the listener to provide a description b_0 of the auditory event. It is this description which helps to provide insight into the character of the auditory event.

The picture may be extended towards a listener in an experiment where quality has to be judged upon. Following Jekosch's definition, the 'quality event' results from a comparison of the perceived composition of the auditory event with its desired composition. Once again, the quality event happens inside the listener; if we want top know about this event, we have to ask the listener for a description. The situation is depicted in Fig. 2, which is taken from Raake [18] and is based on the concepts of Blauert [5][4] and Jekosch [11][10].

The sound event, the auditory event, the quality event and the reference may all be described through their character, comprising each a profile of features [2]. For example, the sound event may be characterized by a set of instrumentally-measurable features (e.g. sound pressure level, equivalent bandwidth and slope of the spectrum, etc.). The auditory event may be characterized by a set of psychoacoustic features, such as loudness, sharpness, roughness, spaciousness, etc. These features are commonly measured with the help of trained expert listeners, in order to reduce inter-individual differences in the feature values.

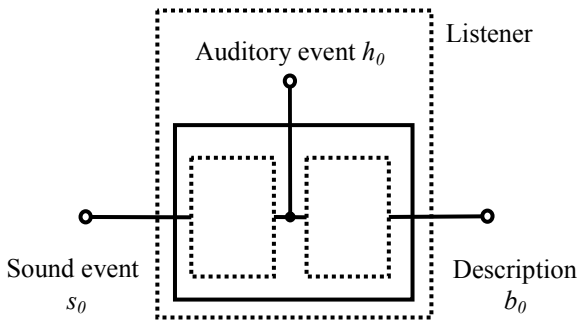

Figure 1: Schematic representation of a participant in an auditory experiment [4].

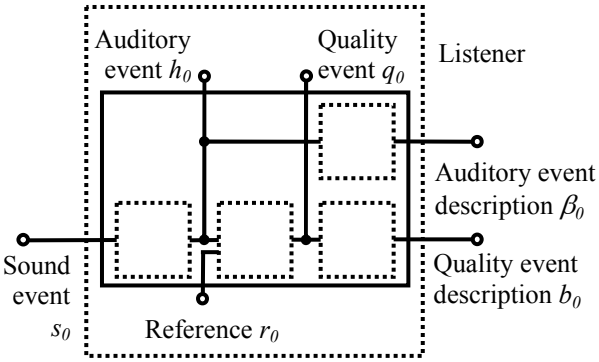

Figure 2: Schematic representation of a participant in a listening quality experiment [18].

The inter-individual differences are assumed to persist in the reference. In fact, the notion of reference comprises all aspects of the judgment process which are determined by the individual listener and its context-of-use, such as individual preferences, abilities, emotions, task specificity, functionality, as well as expectations formed by habituation and tradition. Because of the individuality, the reference can only be assessed with (in general inexperienced and untrained) test participants of the target user group.

Depending on which aspect of quality is to be evaluated, the comparison between the character of the auditory event and the one of the reference takes place on different layers. For the quality related to the sound event, physical features may do the job. When it comes to the quality of the auditory event, psychoacoustic features may be used. However, such features describe the auditory event in a kind of 'neutralized' way, i.e. they do not consider individual preferences, habituation, etc. A comparison of feature profiles which include psychological, semantic and functional factors results in a description of the quality in its context-of-use: The resulting quality aspect may indicate the quality of the system itself, and no longer the quality of the perceptual event. In that case, the evaluation process needs to be carried out with the target participants – not only the reference, but also the weighting of the different features in the judgment process may be highly individual.

So far, we have limited our considerations to passively perceiving (here: listening) test participants. This is a considerable restriction, as many systems acousticians design are interactive ones. Thus, the behaviour of the system – and the sound event associated with it – is largely determined by the input provided by the participant. We therefore have to consider the actions and reactions of the participant as well. For practical purposes, we may assume that our participants are users of a system or service, which have a specific task in mind when they operate the system. Then, they might also have a model of how to perform the task with the help of the system (a so-called 'mental model'). The corresponding task and interaction models will determine the behaviour of the user towards the system. The determination is however not absolute: the user behaviour is largely influenced by the immediate reactions of the system as well, e.g. by the options offered at each instance of the interaction, the sounds emitted, the language used, etc. The user's behaviour can be translated into actions taken towards the system. Whereas the internal behaviour of the user (what the user would like to do) is hidden from the observer, the actions are reflected on an observable surface form. Thus, it is possible to quantify them, and to draw conclusions for quality on that basis. The actions will provoke reactions from the system, and the circle continues anew.

Having identified the processes involved in an interaction experiment designed for measuring quality, these processes may be used in two different ways: First, they provide us hints on how to design the experiment in order to have optimally valid and reliable indices of the construct to be measured. For example, measurement of the sound quality requires psychoacoustic, perhaps also psychological and semantic features to be determined. For the former, trained expert participants may be useful, whereas the latter are better assessed with untrained participants of the target group the measurement result should be valid for. Measurement of service or product quality, in turn, requires a representative context the measurement process can take place in, in order to provoke the right reference inside the user.

Secondly, the processes may help developing algorithms which are able to predict the result of a quality judgment process – without actually carrying out the respective experiment. It may be assumed that such an algorithm reflecting the human perception and judgment processes is able to generalize to unknown physical events and situations. In the following section, the components of a general model for predicting the quality of an interactive service will be briefly defined, following the assumptions about the quality-formation process described above.

3 Components of a model for predicting quality events

Ideally, a general model for predicting quality in an interactive situation would recapitulate the processes happening inside a test participant as closely as possible. It would process the physical signals reaching the perception organs (here: the sound event) into a perceptual event, compare this perceptual event with a reference, and provide a description of the resulting quality event. In addition, the model would be able to act in the way a real user would act, following pre-defined goals, and the user actions would influence the course of the interaction between user and system. While admitting that such a complete model will be far out-of-reach for the moment, we will define some requirements for the individual components in the following paragraphs. The entire picture of all components in given in Fig. 3.

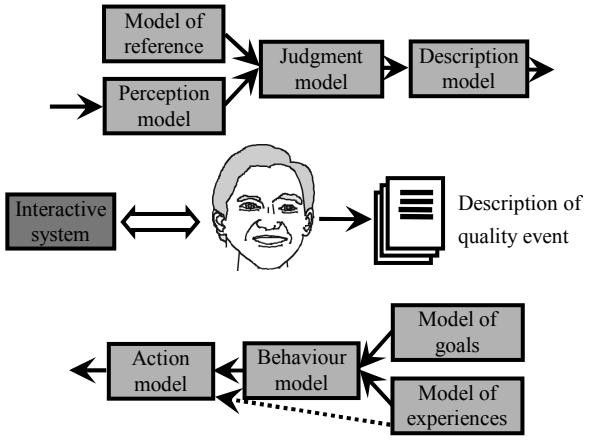

Figure 3: Components of a model for predicting interaction quality.

Perception model: This model should be able to describe the peripheral and cognitive processes transforming the physical event into a perceptual event. For the auditory perception, it may include transformation characteristics of the outer and middle ear, a non-linear frequency analysis, compression laws, models of temporal and spectral masking, etc. In addition, such a model could be able to transform the pre-processed signals into psychoacoustic parameters, such as loudness, sharpness, roughness, noisiness, etc. Assuming normal hearing capabilities, such a perception model would be identical for all possible users, but the character of the resulting perceptual event will depend on the quality judgment task.

Model of the reference: This model includes all individual and functional aspects of the quality-formation process, e.g. the individual preferences of the user, habituation, emotions, task specific aspects, etc. In order to include such a variety of aspects, references may be defined on different levels: On a psychoacoustic level for a direct comparison with the psychoacoustic character of the sound event, when the quality of the auditory event as such is to be measured; on a physical level for a comparison with the character of the sound event, e.g. when it comes to speech-transmission quality; on a psychological and/or cognitive level when the quality of a system or service and its impression on the user are to be measured. It has to be noted that the reference does not need to be stable in the long term; in particular with quickly-changing services it is hard to develop stable references.

Judgment model: The judgment model determines a multi-dimensional distance between the dimensions (features) of the perceptual event and the ones of the reference. Each dimension may either be of the form "the-more-the-better", or have an ideal point where an optimal value is reached. The dimensions may be weighted according to individual preferences or to task-specific requirements. For example, the intelligibility of a telephone connection may usually be of subordinate importance, as long as the usual high intelligibility standard is ensured; in turn, it may be highly important in case of high background noise levels, or when communicating in a foreign language. It should be noted that a distance of zero does not necessarily lead to optimum quality, as the standards set by the reference may well be exceeded.

Description model: This model transforms the result of the comparison process to a well-defined scale. In quality-rating experiments, absolute category rating scales are frequently used, as they are easy to use also for non-experts, and they provide some anchoring on general 'world knowledge' through the meaning of the category labels. However, such scales often show saturation effects at their extremities.

Model of the goals: In task-oriented interactions, the user's goals will mainly be determined by the capabilities of the system. However, there does not need to be a one-to-one relationship; the user might have goals which are not accomplishable with the help of the system. In addition, the user might have a different task structure than the one foreseen by the system. In these cases, interaction failures may occur which impact the interaction quality [17].

Model of experiences: This model may capture the experiences of the user with the task and the domain, but also personal interaction experiences. For example, the user of a speech-based train booking system may have individual requirements for the train in case that he knows which connections are usually available. If he already operated ticketing machines in a railway station, he may also have a specific interaction structure in mind which may determine his behaviour in the current interaction.

Behaviour model: This model captures the interaction behaviour of the user, as a kind of executable interaction machine. In contrast to the model of the goals, this model does not specify *what* the user would like to do, but *how* she would like to do it. The interaction behaviour may be described as a series of individual steps the user performs until the goal is reached.

Action model: The individual interaction steps may be translated into concrete actions the user is likely to perform. For example, the step "specify the name of the destination" in the above-mentioned speech-based train booking system may be translated into a concrete utterance, e.g. "to Berlin main station". The exact wording, as well as supra-linguistic aspects of the spoken utterance, is influenced by the user's experience, as described in the experience model.

The given descriptions may illustrate how a general model of a user judging quality in an interaction experiment might look like. Although no such full model is yet available, individual components already form part of algorithms which are commonly used for predicting quality. The following section will provide some examples of what is available.

4 Available components

The examples given here are taken from the context of quality modelling for telecommunication services. This is due to the fact that the author's background is mainly in this field; in addition, the telecommunication sector seems to have a huge demand for such models, and the results obtained so far are very promising.

Perception model: Perception models are usually part of signal-based approaches to predict speech transmission quality. The speech signal is commonly analyzed in the spectral domain, and loudness values are calculated for each frequency band. In addition, compression and masking effects are taken into account. Some examples also try to model nerve fire behaviour, but this has not yet provided a sufficient advantage over simpler – less detailed – approaches. Details can be found e.g. in [19].

Model of the reference: The reference is probably the most difficult part to model. Simple models for predicting speech transmission quality use a representation of the clean – untransmitted – signal at the entrance of the transmission channel as a model of the reference. This may be surprising, as the human listener usually does not have this reference available when judging upon quality [1]. Single-ended models for speech transmission quality try to artificially generate such a 'clean, undistorted' reference, e.g. with the help of a vocal-tract analysis and re-synthesis [13].

Approaches have also been made to model the functional aspect of the reference. For example, the E-model, a widely-used model for planning telephone networks, describes the functional advantage of mobile phones of connections to hard-to-reach areas in terms of a quality trade-off: it is expected that about half of the degradation inherently associated with a particular service is ruled out by the functional advantage connected to this service [8]. Although this may be considered more like a rule-of-thumb, it indicates possibilities to take functional aspects into account when service quality is to be predicted.

Judgment model: The judgment model has to perform the comparison between the character of the perceptual event and the character of the reference. In simple signal-based approaches for speech transmission quality prediction, this is usually performed as a distance or similarity calculation between the (modified) loudness representations of the input and the output signal. The same holds true for the single-ended quality prediction models, where the loudness representation of the artificially-generated reference is used instead of the one of the clean input signal.

A more psychoacoustically-motivated approach is the attribute-based measure for speech transmission quality [9][20]. The idea is to decompose the perceptual event into orthogonal perceptual dimensions. For narrow-band and wideband transmitted speech, 4-5 such dimensions seem to be important, namely the directness/frequency content, the continuity, the noisiness, as well as the loudness of the speech signal. Estimators for each of these perceptual dimensions may e.g. be determined from the output (and potentially also the input) signal of the transmission channel. The task of the judgment model is to calculate distances between the observed and the 'ideal' values for each dimension, and to weight the dimensions according to their impact on overall quality.

Description model: The task of the description model is to transform the distance or similarity measure of the judgment model into an interpretable index of 'overall quality'. In speech-transmission quality prediction, this is usually performed with the help of a 3rd-order polynomic or a tanh function. Such functions are able to model the saturation at the scale extremities. However, they have to be calibrated if an absolute level of quality – and not a relative comparison between stimuli – is of interest.

Either the judgment or the description model may also include temporal aspects of the quality judgment process. For example, degradations occurring at the end of a telephone call have proven to show a more negative impact on overall call quality than degradations occurring at the beginning of a call. Such recency or end-effects may be taken into account by a proper time-averaging process, as it has recently been shown e.g. in [21].

Models of goals and behaviour: Modelling interaction behaviour with e.g. speech-based telecommunication services is a relatively new field. Whereas simple models try to 'feed' spoken dialogue systems with pre-recorded utterances selected from a previously-recorded databases, the 'MeMo workbench' [15] tries to generate user interaction behaviour on the basis of a description of the system, and of user characteristics. Here, the behaviour is modelled in terms of a probabilistic state machine. Basic probabilities for the user to follow one of the possible paths through the interaction are modified by rules which increase or decrease the respective probability, based either on characteristics of the system (e.g. the prompts of a spoken dialogue system, or the design of the graphical user interface) or the characteristics of the user (e.g. users unfamiliar with the English language are less likely to click a button with an English label).

Model of experiences: The user characteristics of the MeMo workbench may be summarized in a database of prototypical user experiences. In such a database, users are classified in terms of their (expected) interaction behaviour. Proposals in this respect have been made e.g. in [16].

Action model: Action models are well-established in the psychological literature, e.g. the GOMS model which decomposes a task into individual steps (part of the behaviour model), and then associates execution times to the individual steps. In order for such models to work, it may be helpful to dispose of a description of the system (or of the transmission channel) as well. Knowledge on the system may also be used as a part of certain user model components, e.g. for the behaviour model, or for the model of the reference.

5 Conclusions

To the author's knowledge, no complete model of user interaction behaviour and quality perception has ever been implemented, and a lot of work still needs to be done before such a model might be able to generate realistic quality judgments and interaction behaviour. However, the previous section has shown that several building blocks of such a model are already available. Thus, for specific applications, the aim of a complete model for describing quality judgment processes with the aim of predicting quality does not seem to be too far out-of-reach. Disposing of a certain number of models – for different applications – will enable us to draw comparisons between the individual components of successful approaches. In this way, our still incomplete knowledge of quality perception and judgment processes is likely to increase.

The usefulness of models for describing quality formation processes is beyond doubt. Already now, models predicting the quality of speech and video transmission are used in all phases of system planning, implementation, and operation. Such models are by no means meant to replace subjective tests as the 'true' measures of quality; instead, they may be used in cases subjective tests cannot be run, e.g. because the system is not yet available (transmission network planning), or when the large amount of data prohibits subjective testing (e.g. quality monitoring during system operation).

Whereas physical and psychophysical features for describing the physical and the perceptual event – at least for the

auditory perception – seem to be well-established, the character of the reference needs further investigation. This includes the incorporation of experience, of meaning, and of other psychological and cognitive factors. For other modalities, even the perceptual processes are far less modelled. Additional models for modality fusion and fission will be necessary when it comes to describing multimodal perception and action processes. The full picture of a general model outlined in Section 3 will thus be more of a vision – however a very positive one – for the next decades. Following the first steps made by Blauert and Jekosch, the way towards this vision seems to be very promising.

6 Acknowledgement

The present work is partly based on first ideas described in [14]. It was largely inspired by numerous discussions with Blauert, Jekosch, Raake, and other colleagues at the Institute of Communication Acoustics, Ruhr-University Bochum, and later at Deutsche Telekom Labs, Berlin University of Technology. The author would like to gratefully acknowledge the inspiration taken from these discussions.

References

[1] Berger, J., '*Instrumentelle Verfahren zur Sprachqualitätsschätzung – Modelle auditiver Tests*', Arbeiten über Digitale Signalverarbeitung Vol. 13 (U. Heute, ed.), Shaker Verlag, Aachen (1998)

[2] Blauert, J., Jekosch, U., 'Auditory Quality of Performance Spaces for Music – The Problem of the References'. *Proc. 19th Int. Congress on Acoustics (ICA 2007)*, Madrid (2007)

[3] Blauert, J., Jekosch, U., 'Concepts Behind Sound Quality: Some Basic Considerations'. *Proc. 32rd Int. Congress and Exposition on Noise Control Engineering (Internoise 2003)*, Jeju Island, Seogwipo, Korea (2003)

[4] Blauert, J., '*Spatial Hearing: The Psychophysics of Human Sound Localization*', Revised Edition, The MIT Press, Cambridge MA (1998)

[5] Blauert, J., '*Räumliches Hören*', S. Hirzel, Stuttgart (1974)

[6] EN ISO 9000, '*Quality Management Systems – Fundamentals and Vocabulary (ISO 9000:2000)*', International Organization for Standardization, Geneva (2000)

[7] EN ISO 9000, '*Quality Management Systems – Fundamentals and Vocabulary (ISO 9000:2005)*', International Organization for Standardization, Geneva (2005)

[8] ETSI Technical Report ETR 250, '*Transmission and Multiplexing (TM); Speech Communication Quality from Mouth to Ear for 3,1 kHz Handset Telephony Across Networks*', European Telecommunications Standards Institute, Sophia Antipolis (1996)

[9] Heute, U., Möller, S., Raake, A., Scholz, K., Wältermann, M., 'Integral and Diagnostic Speech-Quality Measurement: State of the Art, Problems, and New Approaches'. *Proc. 4th European Congress on Acoustics (Forum Acusticum Budapest 2005)*, Budapest, 1695-1700 (2005)

[10] Jekosch, U., '*Voice and Speech Quality Perception. Assessment and Evaluation*', Springer, Berlin (2005)

[11] Jekosch, U., 'Basic Concepts and Terms of "Quality", Reconsidered in the Context of Product-Sound Quality', *Acta Acustica united with Acustica* 90, 999-1006 (2004)

[12] Jekosch, U., ‚*Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*', Habilitation thesis (unpublished), Universität/Gesamthochschule, Essen (2000)

[13] Malfait, L., Berger, J., Kastner, M., 'P.563 – The ITU-T Standard for Single-ended Speech Quality Assessment'. *IEEE Trans. Audio, Speech, Lang. Process.* 14, 1924–1934 (2006)

[14] Möller, S., Naumann, A., Schleicher, R., ‚Qualitätsplanung und -überwachung interaktiver Telekommunikationsdienste'. *Prospektive Gestaltung von Mensch-Technik-Interaktion* (M. Rötting, G. Wozny, A. Klostermann and J. Huss, eds.), Fortschritt-Berichte VDI Reihe 22 Nr. 25, 407-416 (2007)

[15] Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., Reithinger, N., 'MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations'. *Proc. 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP)*, Pittsburgh PA, 1786-1789 (2006)

[16] Naumann, A., Hermann, F., Niedermann, I., Peissner, M. Henke, K., ‚Interindividuelle Unterschiede in der Interaktion mit Informations- und Kommunikationstechnologie'. *Mensch und Computer 2007*, Oldenbourg Wissenschaftsverlag, München (2008)

[17] Oulasvirta, A., Möller, S., Engelbrecht, K., Jameson, A., 'The Relationship of User Errors to Perceived Usability of a Spoken Dialogue System'. *Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, Berlin, 61-67 (2006)

[18] Raake, A., '*Speech Quality of VoIP: Assessment and Prediction*', John Wiley & Sons Ltd., Chichester, West Sussex (2006)

[19] Rix, A.W., Beerends, J.G., Kim, D.-S., Kroon, P., Ghitza, O., 'Objective assessment of speech and audio quality – Technology and applications'. *IEEE Trans. Audio, Speech, Lang. Process.* 14, 1890–1901 (2006)

[20] Wältermann, M., Raake, A., Möller, S., 'Perceptual Dimensions of Wideband-transmitted Speech'. *Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, Berlin, 103-108 (2006)

[21] Weiss, B., Möller, S., Berger, J., ‚Wahrgenommene Sprachqualität in Telefongesprächen bei zeitlich variierenden Übertragungseigenschaften'. *Elektronische Sprachsignalverarbeitung. Tagungsband der 18. Konferenz* (K. Fellbaum, ed.), TUDpress, Dresden, 210-217 (2007)