Forum Acusticum 2011 - Aalborg - Denmark

About the nature of references – and implications for quality prediction

Sebastian Möller, Marcel Wältermann
Quality and Usability Lab, Deutsche Telekom Labs, TU Berlin, Germany.

Alexander Raake
Assessment of IP-based Applications Lab, Deutsche Telekom Labs, TU Berlin, Germany.

Summary
When judging the quality of auditory stimuli, a listener compares the perceptual event with some internal reference. Up to now, little is known about the exact nature of this reference. In the present paper, we try to identify factors which carry an influence on the internal reference, resulting from the listening environment, the task, the situation in which the perception and judgment take place, the context of use, as well as user-related factors. We further discuss implications from each factor for the design of psychoacoustic experiments which are necessary for measuring quality. Finally, we draw some conclusions for the instrumental prediction of quality, with a focus on speech quality prediction.

PACS no. 43.66.Ba, 43.71.An

1. Introduction

Quality has been defined by Jekosch as the result of a perception and judgment process [1], during which the perceiving human compares the characteristics of the perceptual event with the desired characteristics of some internal reference. So far, little is known about the characteristics of the internal reference. Blauert and Jekosch [2] distinguish e.g. between descriptions on the physical, psychophysical, psychological, and semantic level. In addition, it is assumed that the notion of reference comprises all aspects of the judgment process which are determined by the perceiving human and its context-of-use, such as individual preferences, emotions, task specificity, functionality, as well as expectations formed by habituation.

In order to formalize the quality formation process and make it accessible e.g. to instrumental quality prediction, we propose to decompose the reference into possibly orthogonal features. Such a decomposition is frequently performed when identifying the perceptual space related to auditory stimuli, and methods for the decomposition include similarity or distance scaling with subsequent multidimensional scaling (MDS; McDermott [3]), or semantic differential scaling with subsequent factor analysis (Osgood et al. [4]). We assume that there is a fixed set of features not only for the perceptual event, but also for the reference. However, the features are not always easy to measure, and they are affected by modifying factors. We assume such modifying factors to include environmental factors (physical environment), task-related factors (motivation), situational factors (emotions, distraction), contextual factors (e.g. costs), as well as user factors (individual preferences, habituation). These factors can be assumed as weightings reflecting the importance of perceptual and reference features for the comparison process, the outcome of which is the quality event.

In the following section, we will draw a picture of the quality formation process organized in this way, and we will discuss the implications for psychoacoustic experiments which are necessary for measuring quality. In Section 3, we discuss some implications for modelling these processes, focussing on the application of speech quality prediction. Some conclusions for future work are drawn in Section 4.

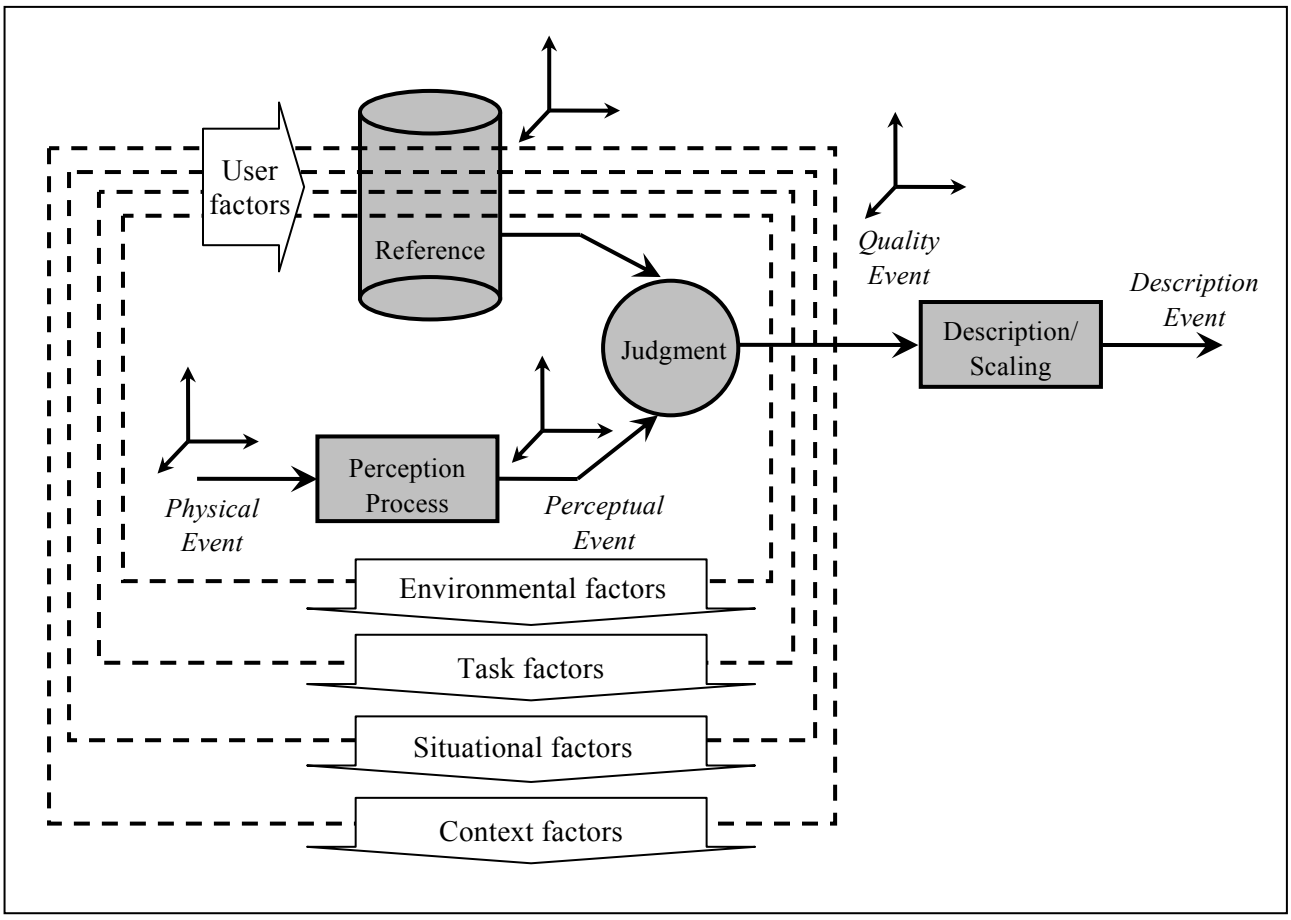

Figure 1 : Processes and modifying factors influencing quality formation in a psychoacoustic experiment.

2. Processes and modifying factors

In the following considerations, we assume that a listener is confronted with auditory stimuli and has to judge their quality. Depending on the listening environment, the task given to the listener, the judgment situation, the context, as well as depending on the user factors (see Section 2.6), the judgment will give rise to different quality descriptions. In the subsequent paragraphs, we first discuss the core events happening inside the listener, and then the modifying factors.

2.1. Sound quality judgment processes

The sound perception and judgment processes are at the core of our interest. For the sake of simplicity, we assume that the perception only takes place in case that we have a physical event, which then provokes a perceptual event happening inside the listener. This perceptual event is compared to the internal task-, context-, and situation-dependent reference.

The physical event is multidimensional, and it can be characterized e.g. in terms of its duration, signal energy, or spectral content. The same holds true for the perceptual event, and – depending on the type of event – dimensions such as loudness, sharpness, roughness, continuity, coloration, or noisiness have been identified [5][7]. Such dimensions have been extracted by either of the above-mentioned methods (similarity scaling and multidimensional analysis, or semantic differential scaling with subsequent factor analysis).

The perceptual event is then compared to an internal reference in order to determine its "value". For the sake of simplicity, we may assume that the internal reference is also multidimensional in nature; this facilitates the comparison process, as a distance can be determined in a multidimensional space. It has to be noted that there are different types of dimensions [6]: So-called "vector-model" dimensions belong to the group "the-more-the-better" or "the-less-the-better"; for these dimensions, the reference values may also be surpassed, resulting in better quality, despite the difference. So-called "ideal-point" dimensions contain an optimum value, and any distance from this optimum is considered as a degradation of the associated quality.

The outcome of this comparison, i.e. the distance between the perceptual event and the reference value, is a quality event, which happens in a particular situation (determined by its place, time, and character). The quality event typically is

multidimensional as well. However, frequently we are interested in a one-dimensional index of quality. In this case, the perceiving human is asked to map this multidimensional event to a one-dimensional scale; obviously, this mapping coincides with a loss of information.

Bare of a specific usage environment, task, situation and context, the quality event and the respective judgment reflects only an abstract, neutralized picture of the quality associated with the perceptual event. Or, more probable, the listener interprets the event as happening in a certain environment and situation, assuming a certain task and usage context, which however is only imaginary – and usually not controlled for in the experimental set-up. The set-up commonly consists of presenting pure stimuli in a quiet listening-only situation of a laboratory. Such a situation is typical for psychoacoustic experiments in which the nature of the sound itself, but not its quality in use, is of interest.

As an example, the quality of speech signals transmitted over standard (narrow-band, wideband and super-wideband) telephone transmission channels has shown quality dimensions such as continuity, noisiness and coloration [7], as long as the level is normalized (which is frequently the case in standard speech quality tests). If the listening level varies, then an additional dimension "loudness" is introduced.

2.2. Environmental factors

Environmental factors affect both, the physical event as well as the reference it is judged against. Here, we are mostly interested in the acoustic factors. For example, background noise may render a low noise level of the to-be-judged acoustic signal imperceptible, or bad loudspeakers and reverberations may mask otherwise notable coloration of an acoustic signal.

On the other hand, environmental conditions have proven to open up new quality dimensions. For example, when listening to speech signals in the presence of background noise, speech quality and noise quality have shown to open up somehow independent quality dimensions. This is reflected in psychoacoustic set-ups of quality judgment, e.g. in terms of separate quality scales for speech and noise in ITU-T Rec. P.835 [8], which then have to be integrated into an "overall quality judgment" by the test participants. So far, it is unknown how the background noise dimension adds to the other quality dimensions observed in transmitted speech. In particular when the background noise occurs at the listening end, the listener might be able to discern two separate sources of degradations, through locally separated cues. Further investigations are necessary to address this issue.

2.3. Task factors

So far, we assumed that the listener's task only consists in listening to the acoustic stimuli and judging their quality. However, this is a fairly unrealistic task; normally, listeners listen to acoustic stimuli because the stimuli are meant to carry information, either in a primary task (speech stimuli for communication, auditory icons and earcons for device feedback, music stimuli for distraction, etc.) or in a secondary, parallel task (e.g. engine noise while driving).

As soon as such a task is defined, the quality judgment process is shaped according to the task characteristics. For example, the task might distract the listener from a very analytic listening process, and subtle details get lost. On the other hand, the task may emphasize certain quality dimensions which would be less important in a neutralized listening situation. E.g. the intelligibility of speech stimuli may get important in case that actual information has to be conveyed, in particular in background noise conditions or when transmission errors occur.

Test procedures have to reflect such a shift in focus. For example, fake tasks have been casted on listeners listening to synthesized speech samples in ITU-T Rec. P.85 [9], in order to distract their focus of attention from the surface form of the perceptual event towards its content. Similar tasks still need to be developed for other application scenarios, e.g. for audio book readings.

A problem arises when the target task is so demanding that parallel listening is no longer sufficient to sensitively extract relevant quality dimensions. In such a case, performance metrics may accompany direct quality judgments. For example, we currently address the usefulness of EEG activity metrics for measuring the cognitive load resulting from coded acoustic stimuli. Or, artificial tasks with the same characteristics as the target task may be designed to obtain more sensitive metrics, e.g. using semantically unpredictable stimuli in case that intelligibility is a target. So far, it is unknown whether purposely changing the acoustic environment (as it is the case with speech reception threshold tests or isopreference tests by adding noise) will still lead to results which are meaningful for a particular target task.

2.4. Situational factors

Even in the same acoustic environment and with the same task, factors resulting from the listening situation may affect the quality event and the corresponding judgment. As an example, assume that speech stimuli of different quality are presented through a high-fidelity sound system. Depending on the assumed origin of the stimuli, the judgment may differ: If we listen to an audio book, we might expect a pleasant voice, high audio bandwidth, etc.; instead, if we listen to a reporter who tells us about a war situation in some remote country, we might consider narrow-band artificial-sounding stimuli with some background noise as perfectly adequate. Here, the assumed origin of the content influences the reference quality is judged against, as well as the weighting of individual dimensions (e.g. coloration and noisiness are judged as less important for the reporter compared to the audio book).

Reflecting the usage situation in a psychoacoustic experiment is not a simple problem. An obvious consequence is to put the listener in a situation which as far as possible reflects the later usage situation, including the listener's motivation, as a kind of mind-setting. The latter might e.g. be conveyed by presenting a real-life usage scenario to the listener before the actual test. To the authors' knowledge, it is yet unclear whether such an artificial bias influences the listeners' judgments in the desired way.

2.5. Context factors

Auditory stimuli are sometimes representative of a product or a service, and as such they are judged with respect to the characteristics of that particular context. For example, a door-closing sound of a small car might be perfectly adequate; however, the same sound might not be adequate at all for a luxurious limousine car. Or, a user might accept noisy and chopped transmitted speech in case that the service is for free; if however the service is paid for, the same stimuli might not be considered to be of sufficient quality any more.

Assessing contextual factors in a laboratory test is very difficult, if not impossible. Even when presenting the stimuli together with the car as a mind-setting, the laboratory situation will not provoke the same feeling, namely that the listener has spent a large amount of money to listen to that particular sound. Field tests are far more realistic in this respect, although the listening situation is then far less controllable.

2.6. User factors

The internal reference is established in the perceiving human over a long period of time. As a consequence, it is important to select the right test participants for a give purpose. In standard psychoacoustic tests in the laboratory, listeners are commonly selected with respect to their perceptual characteristics (normal hearing or hearing loss of a specified degree), sometimes also with respect to their "expertise" ("novice" or "expert" listeners). However, these two aspects are only a subset of characteristics influencing the internal reference. Summarizing the examples stated above, a listener might also be selected with respect to any expected relationship to the stimuli to-be-presented (musician or amateur music listener, expected emotional response to the stimuli), his socio-economic status, his habituation to the type of stimuli under investigation, as well as his individual preferences. It is obvious that the more user characteristics are to be taken into account, the more difficult it will be to obtain a representative number of users, and the more specific the obtained results will be.

3. Implications for speech quality prediction

In the following, we assume that we want to take the mentioned characteristics into account in a model which is able to predict the perceived quality of transmitted speech stimuli, e.g. in a telephony situation. We will try to develop a hypothetical architecture for such a model which explicitly formalizes the mentioned processes and modifying factors.

The first step is to model the multidimensional perception and comparison process. Assuming an n-dimensional space when judging the stimuli in a neutral laboratory situation, we first have to develop estimators for each dimension separately. As an example for this, Scholz [10] developed estimators for the dimensions continuity, coloration and noisiness, and Côté [11] added a fourth estimator for loudness.

The second step is to weight the dimensions with respect to their impact on the listener's quality judgment. For this purpose, and still assuming a neutral and quiet listening situation, Wältermann et al. [12] proposed a simple linear superposition of the dimensions extracted in auditory tests, which can also be applied to dimension estimates. Depending on the type of dimension, it might be necessary to first transform the estimation, so that their amplitude reflects either their quality ("the-

more-the-better") or their associated degradation ("the-less-the-better"). In case that the dimension possesses an ideal point, unfolding will lead to the desired behavior. We have argued elsewhere that the so-called "transmission rating scale" underlying the E-model, a popular quality prediction model used for planning telephone networks, may be a good candidate for performing the integration of dimensions [13].

The third step is to transform the perceptual quality event to a judgment which might have been given in a psychoacoustic quality-rating experiment. This transformation mainly reflects the scale usage in a particular test situation. A simple S-shaped transformation might do a good job, as it considers the limited scale range available to the listener.

The implications of environmental factors in our assumed quality prediction model are twofold: First, the listening environment may influence what can actually be perceived. Changes in the perception process need to be catered for by the dimension estimators: For example, background noise may mask the perception of discontinuities or coloration, and this may actually be modeled by a spectral noise floor in the corresponding dimension estimators. Secondly, additional dimensions or meta-dimensions may be added as a consequence of the physical listening environment. For example, the mentioned distinction between speech quality and noise quality when listening to noisy speech stimuli may be modeled by first estimating a speech quality indicator via the 4 mentioned dimensions, then a noise quality estimator via a spectral and temporal noise analysis (e.g. using the relative approach, cf. Genuit [14]), and then integrating these two estimations into an overall quality estimate. In this integration, it might be necessary to model any interdependency between the noisiness dimension of the speech stimulus and the background noise quality.

Task factors may be considered via amplification or attenuation of particular dimensions in the integration process. For example, the dimension "intelligibility" might be more important in case that the task is to understand a message, and the concurrent dimension "coloration" may become less important in turn.

Situational factors are more difficult to model, as they do affect neither the perceptual event nor the corresponding aspects of the reference. Instead, they influence the semantic interpretation of the reference. Modeling such an interpretation would require to automatically extract the meaning of the stimuli, and to classify them according to a given taxonomy. As long as such an automatic classification is out of reach, a manual labeling and corresponding adjustment of the associated quality judgment situation can be performed. This is e.g. the case in the E-model which assumes an "advantage of access" when a speech service is established to remote areas which are otherwise hard to reach [15][16]; the "advantage of access" is then added to the quality value which would reflect the ordinary telephone usage situation.

Attempts have been made to consider contextual factors such as costs in the quality judgment process. However, it is still under discussion whether costs can be considered as part of the so-called "Quality of Experience", or whether the price should better be a contributing factor to the acceptance of a product or service. Recent discussions (e.g. Kilkki [17]) distinguish between the "Quality of User Experience" and the "Quality of Customer Experience" to cater for this difference.

User factors have to be considered at different points of our model. Whereas perceptual characteristics might require a modeling at the level of the dimension estimators within the core model, individual preferences or habituation may better be modeled at the level of the integration function, or within the mapping function linking the quality event to the scaled quality index.

4. Conclusions

Our perspective of the reference is a theoretical one, and we cannot proof that the mentioned processes and modifying factors exhaustively capture all aspects which are relevant for fully describing the quality formation process. Still, we consider them as useful in order to come up with both assessment methods and prediction models which are valid for a wide range of applications, usage situations and users.

In order to substantiate our perspective, effort has to be dedicated to establishing appropriate psychoacoustic methods. In these methods, a trade-off has to be made between the ecological validity (in terms of modifying factors covered) and the sensitivity of the methods. As soon as the methods are available and have been applied to quantify some of the mentioned influences, more effort can be spent in an adequate quantitative modeling. We do not expect to be able to establish universal models for every purpose, but we think that adjusting parameters within a model will allow us to better reflect influences of the mentioned modifying factors, in order to avoid

systematic mis-estimations of quality prediction models.

References

[1] U. Jekosch: Voice and Speech Quality Perception. Assessment and Evaluation, Springer, Berlin, 2005.

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

[3] B. J. McDermott: Multidimensional Analyses of Circuit Quality Judgments. J. Acoust. Soc. Am., 45(3), 774–781, 1969.

[4] C. Osgood, G. Suci, P. Tannenbaum: The Measurement of Meaning, University of Illinois Press, Urbana IL, 1957.

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

[6] A. Raake: Speech Quality of VoIP: Assessment and Prediction. Wiley, Chichester, West Sussex, 2006.

[7] M. Wältermann, A. Raake, S. Möller: Quality Dimensions of Narrowband and Wideband Speech Transmission. Acta Acustica united with Acustica 96 (2010), 1090-1103.

[8] ITU-T Rec. P.835: Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm. International Telecommunication Union, Geneva, 2003.

[9] ITU-T Rec. P.85: A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. International Telecommunication Union, Geneva, 1994.

[10] K. Scholz: Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen. Doctoral Dissertation, Christian-Albrechts-Universität, Kiel, 2008.

[11] N. Côté: Integral and Diagnostic Intrusive Prediction of Speech Quality. Doctoral Dissertation, TU Berlin, 2010.

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

[13] S. Möller, A. Raake, M. Wältermann, N. Côté: Towards a Universal Scale for Perceptual Value. Proc. Second International Workshop on Quality of Multimedia Experience (QoMEX'10), June 21-23, Trondheim, 2010.

[14] K. Genuit: Objective evaluation of acoustic-quality based on a relative approach. Proc. Inter-Noise 1996, Liverpool, 1996.

[15] ITU–T Rec. G.107: The E-Model, a Computational Model for Use in Transmission Planning. International Telecommunication Union, Geneva, 2009.

[16] S. Möller: Assessment and Prediction of Speech Quality in Telecommunications. Kluwer Academic Publ., Boston MA, 2000.

[17] K. Kilkki: Quality of Experience in Communications Ecosystem. J. Universal Computer Science 14(5) (2008), 615-624.