Speech quality in modern Telecommunications

Jekosch, Ute
Institute of Communication Acoustics
Ruhr-University Bochum
Universitätsstraße 150
44801 Bochum Germany
Tel: +49 234 32 22496 - Fax: +49 234 32 14165
E-mail: jekosch@ika.ruhr-uni-bochum.de

PACS 43.70

Abstract

In the context of telecommunications one of the most important measuring objects is speech quality. Apart from communicative aspects (e.g., way of speaking, listening and interacting behaviour) speech quality can be affected by a number of system parameters of a modern telecommunication service. These parameters are, for example, related to the transmission line (e.g., coders or packet losses), to terminals (e.g., handsets, head-sets, hands-free terminals) and also to new speech technology devices being coupled to transmission systems and terminals (e.g., speech synthesisers, speech recognisers, dialogue systems). In all these cases the quality of the transmitted speech signal depends on these individual system elements and their integration into a telecommunication service.

Practically, speech quality is the result of a perception and judgement process in a communication context. In order to obtain data on perceived quality, measurements are performed. In principle, these measurements can either be carried out with subjects (auditory measurements) or algorithmically (instrumental measurements). Auditory measurements aim at putting to terms otherwise highly individual and sometimes even anonymous perceptual quality events, and instrumental measurements aim at modelling or predicting these.

It is the aim of this contribution to discuss the state-of-the-art of auditory speech quality measurements in telecommunications in view of instrumental measurements. However, what is being said can be applied to different other fields dealing with acoustic/auditory measuring objects as well (e.g., product sound quality).

Resumen

En el contexto de las telecomunicaciones, uno de los objetos de medida más importantes es la calidad de la palabra. Aparte de los aspectos de la comunicación en sí (por ejemplo, la forma de hablar, escuchar y comportarse interactivamente), la calidad de la palabra puede estar afectada por un número de parámetros del sistema de un servicio moderno de telecomunicaciones. Estos parámetros están, por ejemplo, relacionados con la línea de transmisión (por ejemplo, códigos o pérdidas de grupos), con los terminales (p.ej. micrófonos, auriculares, terminales de manos libres) y también con los nuevos instrumentos de la tecnología de la palabra acoplados a sistemas de transmisión y terminales (p.ej. sintetizadores y reconocedores de la palabra, sistemas de dialogo). En todos estos casos la calidad de la señal de palabra transmitida depende de estos elementos individuales del sistema y de su integración en el servicio de telecomunicaciones.

En la práctica, la calidad de la palabra es el resultado de un proceso de percepción y juicio en un contexto de comunicación. Para obtener datos de la calidad percibida se llevan a cabo mediciones. En principio, estas mediciones pueden llevarse a cabo bien con sujetos (medidas con auditorio de personas) o algorítmicamente (medidas con instrumentos). Las medidas con auditorio tienen por objeto cuantificar

eventos que son por otra parte altamente individuales y a veces de calidad perceptual anónima, y las mediciones instrumentales tienen por objeto modelizarlos o predecirlos.

El objeto de esta conferencia es discutir hacer del estado del arte de las mediciones de la calidad auditiva de la palabra en las telecomunicaciones a la vista de las mediciones instrumentales. Sin embargo, lo que se dice también puede aplicarse a otros campo diferentes que traten de objetos de medición acústica/auditiva (p.ej. calidad sonora de productos)

1. Introduction

Today's telecommunication market is very brisk. Again and again new devices are introduced which either improve the acquainted performance of well-known systems or enlarge their scope. The more basic system components (quality elements) are extended for this purpose, the more complex these systems can become. However, it would be both simplistic and inappropriate to see every element of quality as an independent unit that has to have the highest possible quality if a top quality host unit (e.g., a telecommunication network) is to be created from all such elements of quality put together. The criterion for selecting elements is the degree of their contributing to the optimum functional quality of the target unit, and this functional quality is speech quality and communication quality, respectively.

Accordingly, povided that demands on ergonomics, usability and utility are not violated, system elements are not at all a matter of concern for their users. It is a matter for system design and development. For the users the only question is whether the communicative intention is sufficiently supported by the device or not. Consequently, the aim of system suppliers is to offer a service which ensures best speech communication quality. To reach that goal, system engineering (market research, conceptualisation, specification, design) has to have information on how users judge speech quality – not only with regard to its result but also to its process. And they have to know the relationship between system elements and auditory features. Which modifications of elements in the complex system lead to which audible results? In order to collect data on that, auditory speech quality measurements are performed. However, these measurements are often costly and time-consuming so that there is a strong demand for an instrumental measuring system. Input to this measuring system is physical data of, e.g., the transmission path or specific signal characteristics, whose quantities are measured to compute a speech quality index.

Generally, such an instrumental system is of acceptable quality when there is no significant difference between results of instrumental ratings and auditory judgements. To develop such an instrumental measuring system is not an easy task. The method is the following: an algorithm is drafted, and the performance of this instrumental approach is tested against data related to the listeners. These data form the reference for optimisation. If, after a range of performance tests and optimisation, instrumentally measured data on speech quality agree with the listeners' judgements to a specified degree, the instrumental method in question is considered to be valid - or at least promising. After these optimisation procedures the instrumental system is open to be used, e.g., for system engineering to specify a speech transmission system and to select system elements and components in a controlled way.

What has been said so far suggests that auditory test results are 'the' reference for successful system design, both in telecommunications and instrumental speech quality measurement. However, in this context it is often neglected to question the quality of auditory measurements themselves. Generally we distinguish between spontaneous individual judgements and measurements under controlled conditions, e.g., in a laboratory or in a field test. When looking at judgements in these contexts more closely, we can often observe that laboratory and field tests – although using the same telecommunication service – do not always come to the same results. Individual behaviour is not always in line with judgements under controlled conditions. Which consequences can be drawn with regard to auditory speech quality measurements? Does this inconsistency and apparent unpredictability of the judgement behaviour prove that instrumental processes produce better measurement results because they deal with invariance and not with the spontaneous behaviour of the individual, or does it mean that measurements with subjects are done "correctly" because they are working with the measuring apparatus "man"?

Both views must be refuted as will be shown in this article. However, the brief discussion shows that - from the point of view of defining measuring methodologies and methods - auditory and instrumental measurements are closely interrelated: The task of auditory speech quality measurements is to capture those perceptually relevant features users base their judgements on. Similarly, the task of instrumental speech quality measurements is to make use of those system elements or system states which result in quality (when transmitted speech signals lead to an object of audition) and to capture this dependency algorithmically. The better the influencing features (for auditory measurements) and the constitutive elements (for instrumental measurements) are captured the more valid and reliable the quality of both types of measurements will be.

2. Auditory speech quality measurements: a scientific view

Speech quality measurement – be it instrumental or auditory – is not a purely signal-controlled, deterministic event that will necessarily lead to the same conclusions. An auditory quality judgement that relates to the same speech signal can produce totally different results when the process is repeated. There are various reasons for this. It is not only signal characteristics but it is prerequisites, conditions and general processes of perception performance and the way a judgement is reached that have to be analysed if a speech quality judgement is to be reliable and meaningful. If the processes leading to a judgement are not understood, the judgements on speech quality will be questionable. And instrumental speech quality measurements which are based on these data will be obscure as well.

Understanding the judgement processes provides a sufficient basis for designing the concrete task of judging, where you have not only to be aware of what you are going to measure but also of what is being left out. In order to be able to analyse the causality of some of the processes and to research the background of speech quality judgement, it is necessary to analyse the process of speech perception and judgement under the following aspects:

- how do individuals perceive speech quality features? (methods and forms)
- which aspects of speech quality do they perceive? (object)
- why do they judge them in this way? (explanation)
- how good is their ability to judge? (range and limitations)
- how safe and valid is their judgement? (reliability and validity)
- what is their certainty based on when they pass judgment? (reasoning)
- how representative is their judgement? (general applicability)

The aspects mentioned here mainly apply to the judgement process and the degree of certainty of the judgement made. They are structural aids that describe speech events that have either taken place or are in the process of taking place.

Speech quality experts base their experience and knowledge on analysing and describing these events. One of their aims is to obtain findings on how speech quality judgement can be introduced and initiated in a controlled artificial way (e.g., in a test laboratory or in a field test), without the judgements obtained in this way losing meaningfulness when compared to randomly achieved judgements. One practical criterion in engineering is that the amount of effort involved should be minimized (e.g. with regard to the speech material to be assessed and the number of judging listeners). Speech quality assessment as a scientific field therefore provides the knowledge to describe, structure, construct and execute processes that are essential to speech quality judgement. It uses various processes to do this: apart from observing, describing and analysing natural events, one effective method is to deliberately manipulate identified or suspected factors of influence. In this way assessment scenarios can be constructed in clearly defined and reproducible circumstances in a laboratory to increase knowledge of components and links.

Examining influential factors by means of manipulation is a commonly used method. The focus of such experiments lies primarily not in examining the components and their relationships per se, but concentrates on detecting variables and constants in order to obtain a projection model of speech quality assessment that is supported by data. Using a model to describe what has happened is necessary because natural speech quality assessment processes are very complex. Modelling is possible because, in spite of everything, the processes lead to similar results when similar circumstances are present, and several listeners are questioned. Speech quality measurements drive at functional invariances. The listeners' behaviour cannot be put down to arbitrariness, but it is strategic and subject to a certain systematology. The prerequisite of high quality speech quality measurements is to analyse this systematology by means of a structuralistic approach and to obtain a simplified simulation, i.e. a projection model, which agrees with the findings. The principle used for the modelling assumes that the event of a natural speech quality assessment is a structured whole that can be described as a system. In this system, the significance of the components that constitute the whole are characterized by a holistic structure. This structure is an abstract model which is not subject to variations.

One such model which is applicable for speech quality measurements is introduced in Jekosch 2000. It can be used as the basis of a test design. This model based design has the advantage that the artificial scenario does not have a reproductive character, in other words it is not an extract of the world in which judgements on speech quality naturally occur. Moreover, the link between the natural and model based assessment scenario is the analogous categorial structure. It should be pointed out that the analogy is based on the categorial structure, which does not mean that natural and artificial assessment scenarios are structured isomorphically. They cannot be isomorphically structured because in each of the scenarios in which speech quality occurs speech acts have a different communicative function. And this functional aspect has to be captured.

In Jekosch 2000 the field of auditory speech quality judgement is examined as a whole structural concept. In this concept structure is understood as a network of relationships. Therefore the field is divided into components and their interconnecting relationships. With regard to an integral structural concept of auditory speech quality measurement, components are ordered in such a way that any change to an individual component will result in a change to all the other components that are associated with this one. With the help of such a projection model, the whole domain of speech quality assessment is to be understood as a system, driving at invariances. In turn, the system can be used as the reference for the design of both auditory and instrumental measurements.

At first sight there appears to be a contradiction between the method chosen (structural description) and the domain under investigation (speech quality). The basic idea behind the structure initially implies something static, while speech quality is a dynamic event. In other words: The domain has a dynamic character, but the dynamic element is neither a whim nor infinite. The projection model attempts to analyse this dynamism systematically. Thus speech quality assessment is understood as a time-variant system.

In summary: speech quality assessment can be captured systematically by means of a projection model, whereby this model consists of the components and relationships between the individual components. This produces a dynamic network which can analyse variants in time and position – auditorily and instrumentally. The model can be described as a higher level approach where not the methods themselves but the methodologies are formalised.

3. Speech quality measurements: an application oriented view

The question of interest that will be discussed now is related to the reference (namely auditory test results) telecommunication experts base their decisions on when they develop and use instrumental measuring systems. One of the best instrumental measuring system in the area of interconnected network planning (or better: an instrumental tool for network planning) at present is the so-called E-model. It is standardized as Recommendation G.107 of ITU-T. The E-model provides voice-transmission-quality ratings for handset telephony in the standard frequency band for telephone speech, namely, 300 Hz to 3400 Hz. The model provides estimates of the voice transmission quality as perceived at the receivers' side. As it stands, it is based on 18 individual parameters all of which are related to transmission impairment. The computation renders a 'Transmission Rating Factor' R, which is taken as an estimate for voice transmission quality. The usual range of R extends from 0 to 100, where R=100 indicates high, R=0 indicates poor transmission quality. The values can be transformed into corresponding estimates of statistically-oriented quality measures such as

- The percentage of users who classify a connection as good or better, GOB or poor or worse POW. GOB and POW can be measured directly through user interviews directed to the users' opinions on voice transmission quality.
- The percentage of those calls in which intolerable bad transmission leads to an early termination, TME. In general, TME comes out to be smaller than POW.
- The 'Mean Opinion Score' MOS, Measured values of MOS can be obtained from auditory tests with human subjects under controlled conditions.

The intention here is not to go into further details of the E-model (for further details see ITU-T Rec. G.107, 2000, Möller/Raake 2002) but more to discuss the reference of auditory test results the algorithm is based on.

Surprisingly enough, on the instrumental side a transmission rating factor R is computed which ranges between 0 and 100, whereas for auditory assessment data are obtained by averaging mostly on a 5-point absolute category rating scale as, e.g., defined in ITU-T Rec. P.800, 1996. Generally, auditory test data are obtained from articulation and intelligibility tests, from listening only tests using absolute category rating, from listening only tests using paired comparison techniques, from multidimensional analyses and polarity profiles, from talking and listening tests, conversation tests, performance tests, user surveys and usability evaluation. However, scaling and statistical analysis is still a weak point in most of the databases taken as a reference for the E-model so far. Bronwen/McManus state already in 1984 that category scales are often used wrongly, because when analysing the results many assume that they are looking at numerical relatives scaled in intervals which will permit parametrical tests to be carried out. Bronwen/McManus have examined and analysed different category scales. Subjects were given the task of putting levels of known category test scales into a rising sequence. The test was done both for American-English and Italian. It has been shown that category scales do not possess interval properties, and that therefore parametrical statistical tests (which require normal distributions) are not appropriate. Generally, between-point judgements are not allowed (i.e. no 2.5 or 4.5 answers), and observers avoid the end points or boundaries of scales, further reducing the amount of information which can be gathered. These results clearly show that there is very little sensitivity built into these scales. It follows then, that if information other than just rank is sought, more sensitive scaling methods should be used. Such measurement methods such as ratio or graphic scales yield information about distances between stimuli which other measurement methods do not. (Bronwen/McManus 1984:1171)

In other words: To capture and communicate what subjects have auditorily perceived when they listen to a speech event, nothing more than ranking scales are very often used, and these are of low sensitivity. Thereby useful information gets lost. When taking up the point that subjects even avoid the end points or boundaries of scales, in the end speech quality is scaled on a three-point-scale. As a consequence, there is proof that auditory measurements which are taken as a reference for instrumental measurements are extremely superficial or doubtful – although much more sophisticated approaches were available already since 1987 (e.g. category rating scale by Borg/Borg 1987:7).

In spite of this, the E-model performs provably well for traditional handset telephony – even though with relatively low sensitivity. However, if the scope is to be extended to new types of impairments as, e.g., hands-free-terminals, voice over internet protocol or broadband transmission, the database is of no basic use any more because there is a severe lack of understanding the user's behaviour (which is true even for the E-model's current scope of application). Also, so far we have talked about speech quality, but, in fact, what we actually addressed is voice quality. An extension to speech quality means a big step where content and communication aspects will have to be viewed at as well. And here again, understanding judgements is extremely low so far.

Along this line of thinking, experts on scaling heavily criticise instrumental measuring approaches. Borg/Staufenbiel 1992:217, e.g., consider them to offer obscure perspectives in which it is only important to formulate calculations and predictions according to given laws and using constants from collections of formulas, whereby understanding the formulas is secondary if only the predictions are correct. According to their view it is fatal that it is ultimately neither completely clear which field a scale value or index is illustrating, nor under which circumstances this index has to function. As a consequence, the external validation can not change this fact, especially as the criterion is mostly just as vague. Nevertheless, although they are extremely critical, they also see something positive in the approaches: "Despite our main doubt, scaling, as an index, is irreplaceable […] because we can hardly wait for basic research to provide us with something more differentiated. What is more, it is feasible that successfully functioning indices can give access to greater knowledge." (Borg/Staufenbiel 1992:218) In a way this is also true for auditory measurements.

4. Summary

This paper addresses the necessity of having a scientific dispute on the basics of measuring speech quality. The focal point of interest, but also of criticism, does not concern instrumental measurements such as the E-model, but more auditory speech quality measurements. Results of auditory measurements as described above were, almost without exception, used to develop and optimise the E-model. As a consequence, the model's sensitivity and selectivity is comparably low. The aim was to make a contribution to basic research and thus to offer something differentiated to speech quality measurements generally, and to instrumental and auditory measurements in telecommunications specifically. However, what has been discussed here can be applied to different other measuring objects as well, e.g., sound quality.

5. Acknowledgement

Work has been performed at the Institute of Communication Acoustics, RUB (Prof. J. Blauert). The author would like to thank J. Blauert, and the members of her research group (S. Möller, A. Raake, S. Rehmann and S. Schaden) for their support.

6. References

Borg, I., Staufenbiel, Th. (1993), Theorien und Methoden der Skalierung. Bern: Huber

Borg, G., Borg, P. (1987), "On the Relations between Category Scales and Ratio Scales and a Method for Scale Transformation." in Reports from the Department of Psychology, Stockholm University, No. 672. Stockholm: Dept. of Psychology, p. 1-14

Bronwen, L. J., McManus, P. R. (1986), "Graphic Scaling of Qualitative Terms." in Society of Motion Picture and Television Engineers SMPTE Journal, Nov. 1986, Vol 95, p. 1166-1171

ITU-T Darft Rec. P.833 (2000), Methodology for derivation of equipment impairment factors from subjective listening-only tests. Publ. as TD.018 (GEN), ITU-T SG12, May 9-18, International Telecommunication Union, CH-Geneva

ITU-T Rec. P.800 (1996), Methods for subjective Determination of Transmission Quality. International Telecommunication Union, CH-Geneva.

ITU-T Rec. G.107 (2000),The E-Model, a Computational Model for Use in Transmission Planning. International Telecommunication Union, CH-Geneva.

Jekosch, U. (2000), Sprache hören und beurteilen. Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung. Habilitation thesis, Essen University, Germany

Jekosch, U. (2001), "Sprachqualitätsmessungen und Semiotik: Ein interdisziplinärer Brückenschlag." in Fortschritte der Akustik - DAGA 2001, DPG-GmbH, Bad Honnef, 10 p

Möller, S. (2000). Assessment and Prediction of Speech Quality in Telecommunications. Kluwer Academic Publishers, Boston-USA.

Möller, S., Raake, A. (2002). "Telephone Speech Quality Prediction: Towards Network Planning and Monitoring Models for Modern Network Scenarios." accepted paper for Speech Communication.