The future of speech quality evaluation in the presence of AI based implementations

Hans W. Gierlich

Communication Speech and Audio, 52072 Aachen, E-Mail:consulting.gierlich@posteo.eu

Abstract

The evaluation of speech quality is a multidimensional task. Evaluation criteria have to take into account the various conversational situations: listening, talking and interactive conversation, they have to consider the acoustical situation where the talker and listener are in. This concerns background noise and acoustical room conditions. In consequence different types of test signals and different types of test scenarios were developed. The goal is always to define test procedures leading to meaningful, reliable, reproducible results.

Developers of algorithms and terminals always try to improve the system performance by implementing the latest, state of the art technology to improve speech quality in all conversational situations. This stimulates the implementation of techniques automatically adapting to different acoustical situations. While until recently such techniques were just adaptive new implementations use self-learning AI based solutions.

This may change testing strategies significantly. Self-learning algorithms will behave differently over time and ideally, they will improve their performance for their lifetime. However, there is the possibility that learning strategies may not enhance the user's perception in all situations.

This contribution discusses new test strategies, the question, whether the concept of minimum performance requirements may still hold for such systems and which potential alternatives could be seen.

Basic Considerations when Testing Speech Quality

In contrast to the evaluation of any sound or noise, which is only performed in the listening situation, the evaluation of speech quality has to take into account the different conversational situations the user may be in (see **Fig. 1:** Overview of Parameters influencing the perception of speech quality**Fig. 1**).

In the listening situation the focus of the listener is mainly on

- speech intelligibility,
- speech quality,
- And listening effort.

When talking, the focus is on

- Talking effort,
- Echo impairments (generated at the far end).

Perceptually different as well is the conversational situation. Here we can distinguish between conversational effort as a general description and double talk performance as a specific parameter. Conversational effort can be high e.g., due to high transmission delays leading to non-fluent conversation but also e.g., due to impaired or corrupted speech when talking.

Furthermore, the environments the listener or the talker are in, have to be taken into account. On the talker's side (Sending) this maybe different background noise conditions, this maybe room reverberation if the call is performed in a room or this maybe environmental conditions such as wind or rain when the call is performed outside.

Finally, the network has a significant contribution to the perceived speech quality. The network may impair the transmitted speech signal by coding, by transmission errors and by delaying the speech signal. It needs to be noted, that generally we se multiple networks involved, even at the terminal side when a wireless connection is used to connect a hands-free terminal to the actual mobile terminal.

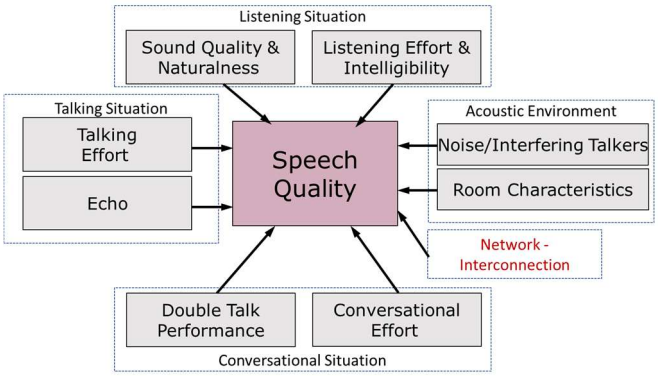

Fig. 1: *Overview of Parameters influencing the perception of speech quality*

When looking forward to more adaptive and AI-based solutions in speech processing we can expect a higher degree of adaptation to different languages and individual talking characteristics (see **Fig. 2**). This can be expected especially in noise cancellation or other type of voice enhancement. When integrating a higher degree of adaptivity in devices an improved noise and/or reverberation cancellation may be possible while better preserving the quality of the individual talker's voice. It is known that speech quality perception is also depending on culture and languages. E.g., the same type of stimuli in identical background situation are rated differently with native listeners in different languages (see [1]). This could be detected automatically and the terminal signal processing could adapt to these features seamlessly. Even an

individual preference setting in a terminal is possible. In consequence, this would lead to a different speech quality rating.

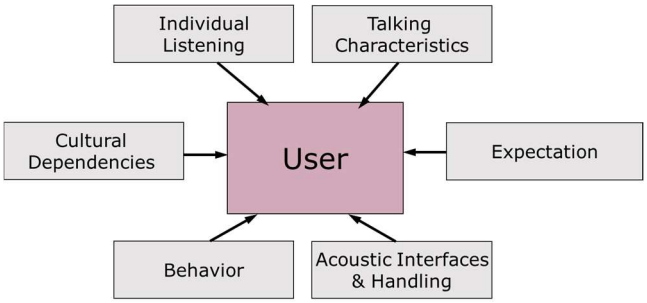

Fig. 2: Speech production and quality perception by individuals

Testing Principles and New Challenges

The general procedures how to reliably test and optimize speech terminals or any software client used for speech communication are targeted to provide a test environment allowing to realistically test in a laboratory type of environment. Therefore, in general the following elements are needed:

- Definition of use cases (handset, hands-free, headset…)
- Analyzing and simulation of the user behavior
- Analyzing the use of the typical environments the terminal is used in and defining appropriate test environments (anechoic, semi-anechoic, office..)
- Defining test sequences
- Defining and simulation of acoustical environments (rooms, background noise …)

These definitions have to be made regardless which type of signal processing is used in the terminals. It is obvious that depending on the type of application these definitions may vary to accommodate the typical use f the different implementations.

If it comes to standardized testing the following principles need to be followed:

„Black-box“ approach: The internal processing and behavior of the device under test (DuT) is unknown. The interfacing to any terminal must be clearly defined and standardized, the test signal and test procedures must be chosen in a way that meaningful results are achieved.

Reproducibility: Test setup, test environment, test signals and analysis procedures must be chosen such that test results can be reproduced within certain accuracy at different locations (e.g. in different labs).

Repeatability: Repeating the tests with all types of terminals within the scope of the relevant test description must lead to the same results within a certain accuracy as given in the test description.

Comparability: Test results of different DuT’s must allow a meaningful comparison of the devices and allow a classification of different devices based on such test results.

The general testing principles not only require clear definitions and procedures for the test environment, the test setup and the test execution, they also assume that terminals behave predictable when exposed to the well described test setup. In former times terminals could be regarded as linear and time-invariant (LTI-)systems. However, adaptive technologies such es speech enhancement, noise- and echo-cancellation are available since quite some time and therefore the LTI-assumption could not be kept since already some time. In consequence the test setups and procedures were changed such that after a conditioning process the DuTs are measured. This assumes that after a conditioning sequence all adaptive algorithms are in a kind of steady state condition and for the time of testing the DuT behaves like an LTI-system concerning the parameter tested. Even with a very limited set of use cases this approach leads to a huge amount of test cases each targeted to investigate a special parameter. The complete test of a mobile terminal according to 3GPP TS 26.131 [2] and TS 26.131 [3] consists of more than 150 testcases depending on which operational modes are tested.

Since a while already and especially with AI-based speech signal processing this testing approach may be questionable. E.g., taking and AI-based noise cancellation we may see a different behavior of the DuT from test to test or even within a test. The DuT captures the desired speech signal as well as background noise. AI models, typically trained on large datasets, are used to distinguish between the desired speech sounds and unwanted background noise. These models can identify patterns to classify speech and noise. In case the AI system identifies noise, an inverse waveform is created to cancel out the noise. In difference to other techniques continuous learning is realized by a feedback loop which enables the DuT to adapt based on new noise profiles. This allows to improve its performance over time and handle different types of noise in various environments more effectively.

When taking this example and applying existing testing technologies it gets clear that current testing procedures would not lead to repeatable and comparable results. For the test of speech in noise a specific, well defined speech test sequence is applied in conjunction with predefined background noises. Applying the identical test in series the DuT may improve its performance from test to test. The speech quality might be preserved better from test to test while the background noise is cancelled out more effectively. So, what would be the test result finally and how meaningful would it be? What should be the requirement for such systems? How would the DuT perform if different speech samples and background noises are applied?

The basic principles for standardized testing as laid out beforehand don’t work anymore when applying the standardized tests as they are used today (e.g. [2], [3])

Possible Ways Forward – Some Basic Ideas

To fully evaluate and cover the potential of AI-based speech processing the well-defined testing methodologies as of today seem not to be a good way anymore. In general, any predefined use case, any predefined test sequence and any predefined test environment allow the speech processing system to adapt to this very specific test environment and

perform a specific learning. In consequence the measured performance varies from test to test and potentially leads to superior results which may not reflect the real-life performance. Furthermore, any divergence from optimum or system errors may not be covered.

So, there is the need to bridge the gap between defining user scenarios and test cases in a standard adequate way on the one hand side and to capture the system behavior even when DuT is constantly changing its properties over time and acquire meaningful, realistic and reliable data. Most likely the complete testing and validation concepts in force today need to be drastically changed. It does not seem to be possible to define

- A fixed test sequence,
- A fix test room and environmental simulation,
- A fix user position,
- Requirements bast on dedicated test procedures described in detail.

Future proof testing must be much more perception oriented; it must contain much less tests but must simulate the typical environment and user behavior more realistically.

The proposal is to define a type of test framework including various time variant positionings of the DuT and user behavior simulation allowing a lot of flexibility and condensed testing in conjunction with a minimum set of performance requirements.

Use Cases

Traditionally use cases for mobile terminals cover the positioning of the device at the user's ear in a fix, standard position (**Fig. 3**) and a fix, well defined hands-free position (see e.g. [4] and [5]). Both in a few well-defined background noise scenarios. These setups are not sufficient. Users move in rooms, they walk on streets and they position their terminals in very different ways, mostly never in the standard position. Some typical examples are sown in **Fig. 4**.

Fig. 3: *Standard handset position as defined in standards*

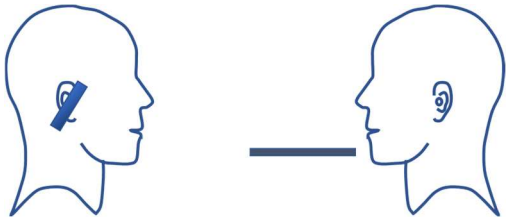

Fig. 4: *Some alternative terminal positionings as used by users today*

This changes the pickup of the speech signal and in consequence the frequency content picked up, but it also changes drastically the SNR conditions as seen by the microphones of the DuT in sending direction. In receiving, similar effects can be observed. The loudness of the transmitted speech signal and the loudness rating, one of the basic parameters in telephonomety, change drastically even in situations with no background noise. In a similar way the sound quality of the speech signal changes due to directivity effects of the microphones and the user changing his position relative to the microphone(s). This effect gets more obvious and annoying the more bandwidth is provided. While in narrowband (300 – 3.4 kHz) with low quality expectations the change of sound quality may be tolerable to some extent; in wideband or super-wideband transmission this is not. It is highly desirable from the user's point of view to compensate the change of speech loudness as well as the varying speech sound quality by adaptive technologies. The same applies for speech in noise. However, here the task is more difficult since in addition to the preservation of the speech sound quality and loudness the noise cancellation needs to adapt to this highly time variant SNR scenario. The terminal noise and echo cancellation performance, the terminals speech enhancement technologies especially in time variant finally determine the user's quality impression and user satisfaction. This can only be evaluated if these changes in user behavior become part of the test procedures and if the variety of room and noise conditions gets higher and more realistic.

In receiving, in general the same issues can be observed: High variation of the loudness, varying sound quality and SNR conditions. When moving the terminal away from the ear, loudness and sound quality decrease mainly to the loss of low frequencies which cannot be produced by the small receivers sufficiently well.

New uses cases and use case scenarios need to be described. These need to include

- user movements,
- varying terminal positions during a call,
- varying background noise scenarios during a call,
- Varying room conditions (reverberation), even during a call.

The simulation of those should be performed in a lab-type environment, providing most realistic simulation covering all these effects.

Test Sequences

A wide variety of test signals are available for telephony testing in ITU-T. Recommendations ITU-T P.50 [6], P.501 [7] and P.502 [8] define test signals and associated testing procedures ranging from simple technical signals such as sine waves up to real speech signals recorded and processed according to well defined rules. When evaluating Non-LTI and/or KI based systems the use of technical or speech like signals in not possible since such systems require real speech signals for proper processing. The speech test signals as defined in Recommendation ITU-T P.501 [7] form a good basis for analysis of these systems, however the currently

defined test signals are not sufficient: More speech signals in all languages of interest need to be available for testing of KI-based systems since the use of the identical test sequence for repeated testing should be avoided. Therefore, instead of using exactly predefined test sequences a framework standard for test signals might be a solution. Such standard might describe the exact way how to acquire the test signals to enable each lab or manufacturer to record their own test sequences. By this procedure pools of test signals can be created which represent the human speech variability in a much better way than fixed sequences defined today. Of course, this leads to a much higher variability in test signals but more representing the real-life situation.

Room and noise conditions

In a similar way the room and background noise conditions may be changed. Instead of using fixed background noise scenarios and room reverberation simulation defined e.g. [9] a framework standard how to record noises and environmental impulse responses may be a way forward to increase the variability of testing.

Analysis Procedures

It is obvious when using a much higher variability or real speech signals in combination with a higher variability of room and noise conditions the huge number of tests used today are no longer applicable. Furthermore, all signal-based tests ideally should be replaced by perceptually based tests, ideally not even using a reference signal but purely using the acquired signal for perceptional analysis. These procedures must be trained to perform well with all types of signals, room and noise combinations and being robust and well performing when new signals are applied. Most likely such procedures would be KI-based as well. The following parameters are proposed to be investigated at minimum:

- "Total" speech quality,
- Conversational quality,
- Conversational efficiency,
- Speech quality in listening and talking,
- Listening and talking effort.

Generic test setups

If a framework approach is used future testing requires a highly automated and easily adaptable test environment. User and terminal movements must be simulated easily and configurable in a laboratory environment. Here again a framework standard could define how such positioning and movement simulation is to be performed. In parallel to the movement simulation of the DuT and the simulated user different speech sequences simulating different conversational situations are applied. Ideally, none of the signals and the signal – noise combination is used repeatedly. For any new test a new sequencing is generated automatically by the system based on the framework standards to be developed. The analysis of the different parameters is performed at the relevant section of the test sequences. A possible example is shown in **Fig. 5**.

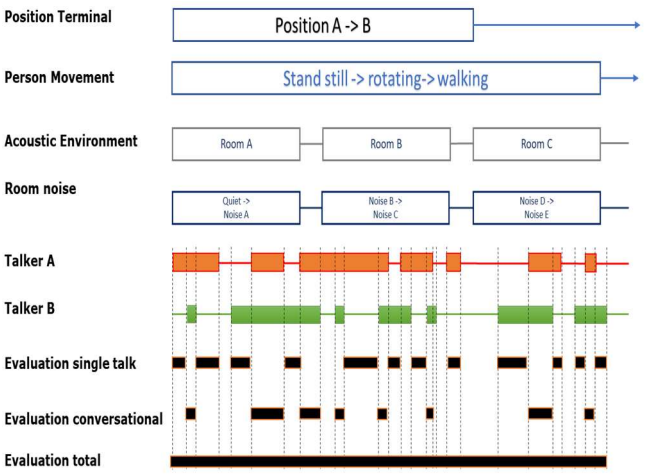

Fig. 5: *example for a possible test sequence and related analysis techniques*

Conclusions

Some ideas were presented how to evaluate Non-LTI and AI based terminals for speech communication in the future. It is obvious that a paradigm change compared to standard test as performed today is necessary. New framework standards need to be developed and it needs to be discussed how performance limits can be established and what accuracy of testing can be expected.

References

[1] ETSI TS 103196: Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods, ETSI 07/2021

[2] 3GPP TS26.131: Universal Mobile Telecommunications System (UMTS); LTE; 5G; Terminal acoustic characteristics for telephony; Requirements, 05/2025

[3] 3GPP TS 26.132: Universal Mobile Telecommunications System (UMTS); LTE; 5G; Speech and video telephony terminal acoustic test specification, 05/2025

[4] Recommendation ITU-T P.64: Determination of sensitivity/frequency characteristics of local telephone systems, ITU-T 11/2007

[5] Recommendation ITU-T P.581: Use of head and torso simulator for hands-free and handset terminal testing, ITU-T 02/2014

[6] Recommendation ITU-T P.50: Artificial voices, ITU-T 09/1999

[7] Recommendation ITU-T P.501: Test signals for use in telephony and other speech-based applications, ITU-T 05/2020

[8] Recommendation ITU-T P.502: Objective test methods for speech communication systems using complex test signals, ITU-T 09/2014

[9] ETSI TS 103 224: A sound field reproduction method for terminal testing including a background noise database, ETSI 03/2020