EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

SPEECH QUALITY EVALUATION OF HANDS-FREE TERMINALS

H. Klaus, E. Diedrich, A. Dehnel, J. Berger
Deutsche Telekom Berkom GmbH
Goslarer Ufer 35, 10589 Berlin, Germany
Tel. (+49 30) 34 97-23 82, Fax: (+49 30) 34 97-29 62, E-mail: H.Klaus@Berkom.De

ABSTRACT

This paper describes a new methodology for the speech quality assessment of hands-free terminals and discusses the results of a pilot study performed in 1996 at the Berlin laboratories for speech quality assessment at the Technology Centre of Deutsche Telekom. Up to now, critical speech quality aspects of hands-free terminals are usually assessed with conversational tests. With the test method proposed here, much more efficient listening only tests can be applied to evaluate various speech quality aspects of hands-free terminals. In the pilot study, a series of conversational tests, specific double talk tests and listening only experiments were performed. The paper descibes the recording environment and equipment, the auditory test methodology and the results of the listening only experiments.

1 INTRODUCTION

Up to now no sufficient test methodologies for the assessment of speech quality of hands-free terminals (HFT) are existing. The main problem is that the critical aspects mainly occur in a conversation between subscribers using HFT's. This was the reason to develop a new methodology for conversational tests described in [1].

Each kind of conversational tests with two partners in a test laboratory is very time consuming and requires a great number of test persons. With the test method proposed here, much more efficient listening only tests can be applied to evaluate various speech quality aspects of hands-free terminals. Since the subjects have only to judge the transmitted speech quality without any loss of concentration for their own speech activities, it is possible to ask for more specific attributes than under conversational test situations. Furthermore, these listening only test procedures described below are only one aspect among a set of different test methodologies (conversational test, double talk test, objective/instrumental measurement techniques) created to establish a unified measurement methodology to evaluate the speech transmission quality of hands-free terminals.

2 METHODOLOGY

2.1 Speech material

The test situation can be described as a strongly controlled simulated conversation produced by two different double talk sequences. To establish this conversation scenario and to record the speech samples, two artificial talking and listening heads were used in a well defined acoustical environment. Each artificial head was located in a separate office room of usual size and acoustical properties. The speech material at the "far end" consists of five short German sentences, fluently spoken by an artificial talking head with a male voice. The received speech material were recorded binaurally using an artificial talking and listening head at the "near end", the observation room. This artificial head presented its speech samples spoken by a female voice, simulating the double talk situation during active phases at the "near end".

Two different scenarios were realized, a **short time** (Fig. 1) and a **long time** (Fig. 2) double talk sequence.

a) „Wir wollen heute spazieren gehen. Zuvor müssen wir uns stärken. Dazu essen wir den Salat. Die Kartoffeln gehören zum Mittagessen. Danach tut eine Wanderung gut.“

b) „Vater hat den Tisch gedeckt. Der Kaffee dampft in den Tassen.“

*Fig. 1: Texts simulating a **long time** double talk sequence. The reading of text b) starts immediately after the second sentence of text a) and occurs while the underlined passage is spoken.*

a) „Wir wollen heute spazieren gehen. Zuvor müssen wir uns stärken. Dazu essen wir den Salat. Die Kartoffeln gehören zum Mittagessen. Danach tut eine Wanderung gut.“

b) „Freilich!“

*Fig. 2: Texts simulating a short **time double** talk sequence. The short response (the German word „freilich“) is spoken at the end of sentence 4 during the German word „Mittagessen“.*

The tests described in this paper contain six real HFT's and in addition a collection of simulated conditions. These offline simulations represent typical effects of connections between HFT's as follows:

- Variable talker echo return loss (12, 16, 20 and 24 dB);
- Level difference between single talk and double talk (2, 4, 8 and 12 dB);
- Completeness of speech transmission (4 different intensities);
- Switching characteristics between single talk and double talk (4 different characteristics).

The validity of the listening only tests can be derived by comparison of its test results with results gained by conversational tests. Those conversational tests were performed with the same systems under test.

2.2 Listening Experiments

A series of three listening only experiments was performed in order to evaluate different speech quality aspects. As shown in Table 1, each of the experiments is subdivided into 3 sessions. All speech samples were presented to the subjects with headphones. For experiment 1, a 5 point absolute category rating scale was applied; for experiments 2 and 3, an adapted 5 point degradation category scale were used, both according to ITU-T Rec. P.80.

In total, 23 experienced listeners (Group 1) and 2*24 naive subjects (Groups 2 and 3, respectively) participated in these experiments. The members of Group 1 were scientists with experience in speech communication systems and/or in auditory test methodologies. This group contained two female and 21 male subjects. The Groups 2 and 3 consisted of 24 "normal", naive persons from a pool of test persons used for auditory tests. According to age and educational background, a good approximation of the normal telephone users' population was aspired. Both Groups 2 and 3 consist of 12 female and 12 male subjects each. All subjects were previously audiometrically checked for normal hearing threshold by pure tone audiometry.

In addition to written instructions, a short conversation via two HFT's was demonstrated in the assessment laboratory in order to clarify the handling of the scoring procedure and to illustrate the test situation, i.e.: The subjects are observers ("ear-witnesses") of a conversation via HFT's. To illustrate the test situation as best as possible and to reflect the recording scenario as described in section 2.1, a female staff member within the assessment laboratory and a male far end speaker introduced the subjects. The basic message of the instruction given to the subjects was as follos:

- Assume that you are located behind a phoning woman and listen to the conversation.
- You have to assess only the speech quality of the caller party, i.e. the male voice at the far end.

In order to establish an internal individual quality reference scale, each part of the experiments started with a short training sequence of 8 speech samples.

No.	Speech Quality Aspect	No. of subjects	
		exp.	naive
1	**Assessment of overall speech quality and sound impression** a) Overall speech quality (long time double-talk sequences) b) Overall speech quality (short time double-talk sequences) c) Sound impression (single talk only)	Group 1	Group 2
2	**Assessment of different speech impairments** a) Impairments caused by speech gaps b) Impairments caused by echoes c) Impairments caused by loudness variations	Group 1	Group 3
3	**Assessment of impairments with respect to double-talk** a) Impairments caused by loudness variations b) Impairments caused by level differences between single talk and double talk sequences c) Impairments caused by switching characteristics	Group 1	—

Table 1: Listening experiments for evaluation of speech quality aspects of hands-free terminals.

3 Results

A subset of the test results of experiments 1 and 2 are given in Fig. 3 to 8 on the next page. Due to the limited space, only naive listener's test results can be presented here. Each diagram shows the results using bar graphs which represent Mean Opinion Scores (MOS) of the test subjects of the specific listener's group, including information about the confidence interval at a confidence level of $1-\alpha = 0.95$.

Hands-Free Terminals

Experiment 1

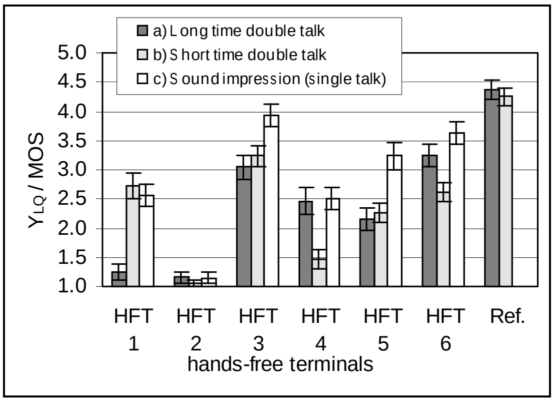

Fig. 3: Overall speech quality assessed by Group 2

Experiment 2

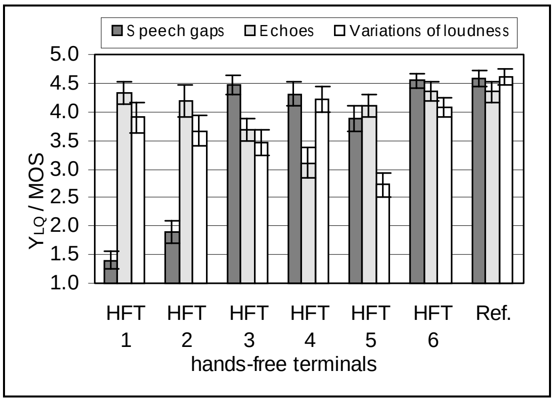

Fig. 4: Speech Impairments assessed by Group 3

Simulated Transmission Scenarios (Experiment 2 only)

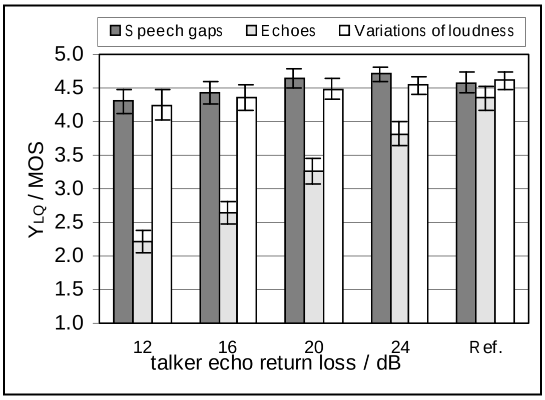

Fig. 5: Echo disturbances caused by talker echo return loss

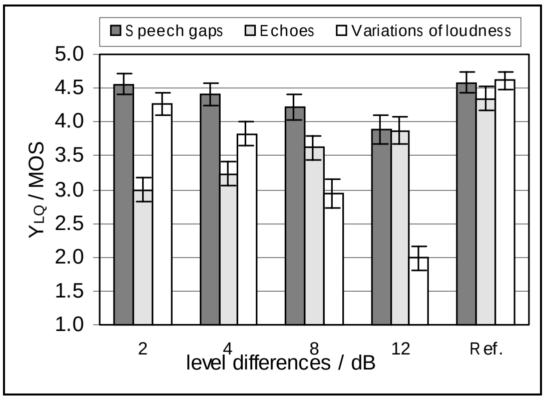

Fig. 6: Level differences between single talk and double talk

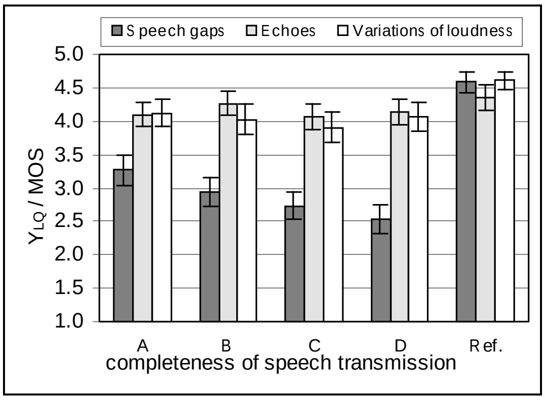

Fig. 7: Completeness of speech transmission

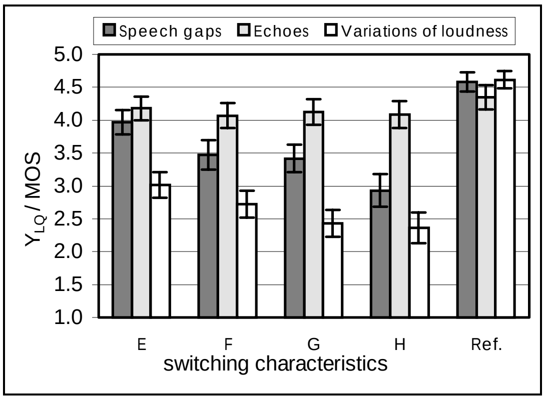

Fig. 8: Switching characteristics between single and double talk

3.1 Experiment 1

The scoring of the real hands-free terminals of Fig. 3 indicates that the subjects were able to differenciate clearly among different terminal qualities. The sizes of the confidence intervals are comparable to codec assessment results and emphazize the power of this test methodology.

The MOS values of additional conditions processed by offline simulations were used for validation and cross-check purposes. They indicate that the experimental design ensures correct values for the perceived speech quality. The assessment results of can be summarized as follows:

- As expected, the long time double talk sequences are much more critical in this experiment than the short time double talk sequences. A decreasing completeness of the transmitted speech information decreased the Mean Opinion Scores of all simulated transmission scenarios. For short time double talk sequences, the differences among the MOS values were not significant.

- As expected, decreasing talker echo return loss values lead to decreasing quality scores. During short time double talk sequences, the echo disturbances were more annoying than during long time double talk sequences. It is remarkable that the experienced subjects judged more critical than the naive subjects.

3.2 Experiment 2

The results in Fig. 4 show that the subjects were able to distinguish clearly between different terminal qualities (as in Experiment 1) and also among different questions. Furthermore, the subjects assessed the variation of characteristics of the HFT's very clearly. For example, HFT 1 and 2 were using level switching components. Here, the Mean Opinion Score for speech gaps is very low. For these terminals, no noticable echo was indicated.

In opposite to Experiment 1, however, experienced subjects (Group 1) were not much more critical than naive test persons (Group 3).

The scoring results of different kinds of speech impairments produced by simulated transmission scenarios are presented in Figures 5 to 8 and can be summarized as follows:

- **Echo disturbances (Fig. 5):** Different scorings were achieved for these speech samples only in Experiment 2b (*middle bar graph, light gray*). The subjects were able to detect echoes correctly. Furthermore, they were able to distinguish clearly between speech gaps and echo effects.

- **Level differences (Fig. 6):** For increasing level differences, the scoring is decreasing very strongly in Experiment 2c (*third bar graph, white*). Furthermore, it can be seen that a slight level difference was perceived as an echo (*middle bar graph, light gray*). A slight but not significant decreasing trend is also detectable in Experiment 2a (*first bar graph, dark gray*).

- **Completeness of speech transmission (Fig. 7):** As expected, a decreasing completeness of the transmitted speech information leads to decreasing Mean Opinion Scores in Experiment 2a (*first bar graph, dark gray*). Here, the experienced subjects (Group 1) were much more critical than the naive subjects (Group 3). For Experiment 2c, however, the occurrence of speech gaps had no influence on the judgements.

- **Switching characteristics (Fig. 8):** Different switching characteristics influenced the scorings for Experiments 2 a and 2c only.

3.3 Experiment 3

This experiment addresses in particular the speech quality impairments during double talk. Only experienced subjects were used because some background knowledge of speech processing principles is necessary to handle this specific assessment task properly. In summary, the listening only test methodology ensures consistant results for the experienced listeners (Group 1) for all experiments.

4 Conclusions

Correlations between listening only and conversational tests as described in [1] indicate that with the new listening only test methodology presented here, specific speech quality impairments of hands-free terminals can be evaluated analytically – in addition to more overall quality aspects gained from conversational and double talk test scenarios [2]. Therefore, a combination of these test procedures is able to specify the achieved speech transmission quality of hands-free terminals in very detail. From the huge amount of data gained by the listening experiments presented here, a number of speech quality parameters will be derived that can also be estimated by instrumental speech quality measurement techniques.

5 References

[1] Subjective evaluation of hands-free telephones using conversational tests, specific double talk tests and listening only tests. ITU-T Contribution COM 12-6-E. Geneva, April 1997.

[2] Subjective evaluation procedures for hands-free telephones – double talk performance. ITU-T Contribution COM 12-5-E. Geneva, April 1997.