



INTERNATIONAL TELECOMMUNICATION UNION

**ITU-T**

TELECOMMUNICATION  
STANDARDIZATION SECTOR  
OF ITU

**P.851**

(11/2003)

SERIES P: TELEPHONE TRANSMISSION QUALITY,  
TELEPHONE INSTALLATIONS, LOCAL LINE  
NETWORKS

Methods for objective and subjective assessment of  
quality

---

**Subjective quality evaluation of telephone  
services based on spoken dialogue systems**

ITU-T Recommendation P.851

---

ITU-T P-SERIES RECOMMENDATIONS

**TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS**

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Subscribers' lines and sets	Series	P.30 P.300
Transmission standards	Series	P.40
Objective measuring apparatus	Series	P.50 P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
<b>Methods for objective and subjective assessment of quality</b>	<b>Series</b>	<b>P.80</b> <b>P.800</b>
Audiovisual quality in multimedia services	Series	P.900

*For further details, please refer to the list of ITU-T Recommendations.*

## **ITU-T Recommendation P.851**

### **Subjective quality evaluation of telephone services based on spoken dialogue systems**

#### **Summary**

This Recommendation describes methods and procedures for conducting subjective evaluation experiments for telephone services which are based on spoken dialogue systems. The respective systems enable a natural interaction via spoken language and possess speech recognition and interpretation, dialogue management, and speech output capabilities. The set-up and running of appropriate interaction experiments is described, and questionnaires for quantifying the relevant quality dimensions perceived by the user are given.

#### **Source**

ITU-T Recommendation P.851 was approved on 13 November 2003 by ITU-T Study Group 12 (2001-2004) under the ITU-T Recommendation A.8 procedure.

#### **Keywords**

Dialogue management, interaction parameter, speech generation, speech recognition, speech understanding, spoken dialogue system, subjective evaluation.

## FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

## NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure e.g. interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

## INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

© ITU 2004

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

## CONTENTS

	<b>Page</b>
1 Scope .....	1
2 References.....	1
3 Abbreviations.....	2
4 Introduction .....	2
4.1 Tasks and components of a spoken dialogue system .....	2
4.2 Telephone interaction with a spoken dialogue system .....	3
4.3 Quality aspects and influencing factors.....	4
4.4 Subjective evaluation methods .....	7
5 Spoken dialogue system characterization.....	8
5.1 Agent factors .....	8
5.2 Task factors .....	10
5.3 User factors.....	11
5.4 Environmental factors .....	11
5.5 Contextual factors.....	11
6 Experimental set-up.....	12
6.1 System set-up and Wizard-of-Oz simulation .....	12
6.2 Test scenarios .....	13
6.3 Test subjects .....	14
7 Questionnaires .....	15
7.1 Questions related to the user's background .....	16
7.2 Questions related to the individual interaction.....	18
7.3 Questions related to the user's overall impression of the system .....	20
8 Usability evaluation.....	22
9 Analysis and interpretation of collected information .....	23
Appendix I – Scenario examples .....	24
BIBLIOGRAPHY .....	26



# ITU-T Recommendation P.851

## Subjective quality evaluation of telephone services based on spoken dialogue systems

### 1 Scope

This Recommendation describes subjective evaluation methods providing information about the quality of telephone services based on spoken dialogue systems, as experienced by the users of such services. Spoken dialogue systems addressed by the Recommendation enable a spoken language interaction with a human user via the telephone network on a turn-by-turn basis, and have speech recognition, speech understanding, dialogue management, response generation, and speech output capabilities. They may provide access to information stored in a database, or allow different types of transactions to be performed.

The evaluation methods described here address different aspects of quality from a user's point of view, taking the spoken dialogue system as a black box. Important quality aspects are the usability of the service, the communication efficiency, task and service efficiency, user satisfaction, perceived speech input and output quality, the system's cooperativity, the symmetry of the interaction, and the perceived smoothness of the interaction. The methods are based on laboratory experiments in which subjects interact with the spoken dialogue system in order to perform a pre-defined, realistic task. The subjects' opinion on perceptive quality dimensions can be rated in a guided or unguided way, on questionnaires that are given to them after the experiment, or with the help of other usability evaluation methods. This Recommendation describes the set-up and running of interaction experiments, relevant quality dimensions perceived by the user, and methodologies that will provide information about these quality dimensions. Further guidance on subjective evaluation methods in general and on the assessment of speech output devices is available in ITU-T Recs P.800 and P.85, and in the Handbook on Telephonometry.

### 2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

- ITU-T Recommendation E.800 (1994), *Terms and definitions related to quality of service and network performance including dependability*.
- ITU-T Recommendation G.107 (2003), *The E-Model, a computational model for use in transmission planning*.
- ITU-T Recommendation G.1000 (2001), *Communications Quality of Service: A framework and definitions*.
- ITU-T Recommendation P.85 (1994), *A method for subjective performance assessment of the quality of speech voice output devices*.
- ITU-T Recommendation P.800 (1996), *Methods for subjective determination of transmission quality*.
- ITU-T Handbook on Telephonometry (1992).

### 3 Abbreviations

This Recommendation uses the following abbreviations:

ACR	Absolute Category Rating
ANOVA	Analysis of Variance
ASR	Automatic Speech Recognition
CCR	Comparison Category Rating
DARPA	Defense Advanced Research Projects Agency
DCR	Degradation Category Rating
DTMF	Dual Tone Multiple Frequency
HMM	Hidden Markov Model
HSD	Honestly Significant Difference
MOS	Mean Opinion Score
MLP	Multi-Layer Perceptron
PARADISE	PARAdigm for DIAlogue System Evaluation
QoS	Quality of Service
SDS	Spoken Dialogue System
WoZ	Wizard-of-Oz

### 4 Introduction

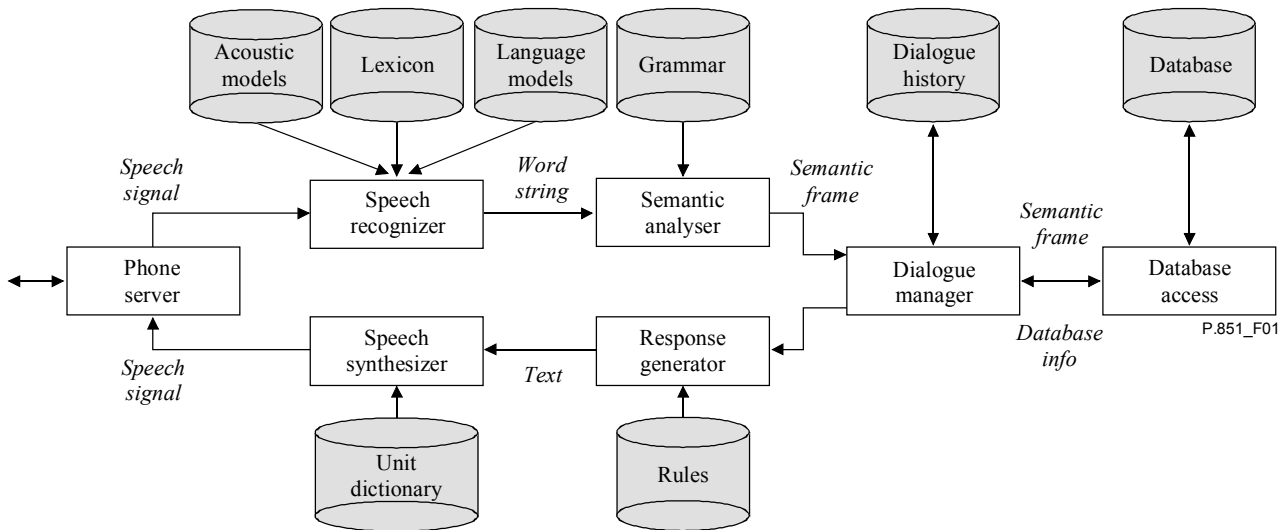
Spoken dialogue systems (SDSs), i.e., computer systems with which human users interact via spoken language on a turn-by-turn basis, may be part of modern telephone networks. They enable access to databases and transactions via the telephone, e.g., for obtaining train or airline timetable information, stock exchange rates, tourist information, or to perform bank account operations, or make hotel reservations, etc. In contrast to simple DTMF systems, spoken dialogue systems possess automatic speech recognition and speech understanding (i.e., syntactic/semantic/pragmatic and thus interpretatory) capabilities, and a dialogue management module that ensures the smooth and natural run of the spoken interaction between the user and the system. As a result, the interaction becomes more human-like, and the service provided by such systems may attract a wider range of potential users. Frequently, DTMF- and spoken-dialogue-based types of systems are implemented in an integrated way, and a part of the respective quality aspects will be identical for both types of systems. Sometimes, spoken-dialogue-based types of systems make use of structures and interface protocols used in web application environments, and are built in a similar way to web interfaces; thus, web interfaces may form a reference for obtaining the same functionality.

#### 4.1 Tasks and components of a spoken dialogue system

From a technical point of view, the components of a spoken dialogue system, operated over the telephone network, can be best displayed in a sequential structure. An example of such a structure is depicted in Figure 1. It consists of six major components which are accessed by the user via a phone server interface. The speech signal from the user is first processed by the speech recognizer. During the recognition process, it is transformed into a word string or a word hypothesis graph which is then semantically analysed. The output is a semantic frame representing what has been "understood" from the user's utterance. It is the task of the dialogue manager to interpret the semantic frame in the context of the dialogue and the task, and to keep track of the dialogue history. When all relevant information has been collected from the user, a query to the underlying



application (in this example a database) can be launched. The information originating from the application program, as well as other communicative goals of the dialogue manager, has to be transformed into a response for the user. This is the task of the response generation module. It generates a response in text form, which is then transformed by the speech synthesizer into a speech signal which is transmitted to the human user. Sometimes, response generation and speech synthesis are implemented as a single module (without stepping to the textual representation), and pre-recorded messages are used instead of synthesized speech.



**Figure 1/P.851 – Sequential structure of a telephone-based spoken dialogue system [27], [33]**

This principle structure may be implemented in different ways. Examples can be found in [4]. One popular way is the so-called "hub architecture" [37], [45] which is used in the DARPA Communicator project. Other structures rely on asynchronously operating modules for interpretation, behaviour (reasoning and acting), and generation; see [1].

## 4.2 Telephone interaction with a spoken dialogue system

The interaction with the spoken dialogue system takes place via some type of telecommunication network. This network will introduce a number of transmission impairments which will impact the quality of the transmitted speech, and as a consequence also the performance of a speech recognizer, and of subsequent speech and natural language technology components in the spoken dialogue system. On its way back to the human user, the transmission channel will degrade the speech signal generated by the dialogue system. Because telecommunication networks will be confronted with human-to-human communication as well as human-machine-interaction scenarios, it is important to consider the requirements of both the human user and the speech technology device. The requirements will obviously differ, because the perceptive features influencing the user's judgement on quality are not identical to the characteristics of a speech technology device, e.g., of an automatic speech recognizer (ASR).

The human user carries out the interaction via some type of user interface, e.g., a telephone handset, a hands-free terminal, or a headset. The acoustic characteristics of the mentioned interfaces are very diverse, and so is their sensitivity to room acoustic phenomena occurring in the talking and listening environment of the user. For example, ambient noise may significantly impact the intelligibility of speech signals transmitted through a hands-free terminal, and it also carries an influence of the talking behaviour of the user. As a result, the whole interaction scenario including the spoken dialogue system, the transmission channel, and the user interface has to be taken into account for the overall quality of the interaction.

### 4.3 Quality aspects and influencing factors

Humans are the users of SDS-based services which are offered over the phone. Thus, human factors have to be taken into account when the functions of a system/service and the degree of their fulfilment are determined. The quality of the service results from the perceptions of its user, in relation to what they expect or desire from the service. Following a definition of quality developed in [24], the *quality of a spoken-dialogue-system-based service* is the result of appraisal of the perceived composition of the service with respect to its desired composition. Thus, the quality perceived by the user is a compromise between what he/she expects or desires, and the characteristics he/she perceives while using the service. It is highly dependent on the situation in which the perception and judgement take place. This fact has to be taken into account when carrying out subjective quality evaluation experiments, namely by creating a more-or-less natural test situation and a realistic test user motivation.

In contrast to the notion of speech transmission quality in human-to-human communication scenarios, the user of a spoken-dialogue-system-based service takes an active part in speech production and dialogue flow. Thus, user characteristics and behaviour may be decisive for the fulfilment of the desired task. The system and service characteristics will therefore be largely influenced by the user. In order to describe the behaviour of the system and the user in a simplified way, parameters can be logged during the interaction. Such interaction parameters may, but need not be instrumentally measurable. Examples include the number of utterances or the duration of a dialogue which are instrumentally measurable, or a word error rate and a task success measure which can only be determined with the help of human experts. An overview of interaction parameters can be found in [18] and [33].

In principle, the quality of a spoken-dialogue-system-based service can be addressed from two different points of view: the one of the service provider and the one of the user<sup>1</sup>. The service provider is mainly interested in the effects of individual elements of the service, and how they relate to the user's degree of satisfaction or acceptability. Service providers make use of the definition of the Quality of Service (QoS) given in ITU-T Rec. E.800. The user perceives and reflects on perceived characteristics (features) of the service, compares the perceptions with some type of internal reference, and judges them according to whether they fulfil his/her expectations or desires. When investigating the quality of a service, it is important to take both points of view into account. Subjective evaluation methods as the ones described in this Recommendation will concentrate on the user's point of view. They will, however, be useful for the service provider as well, as they give indications about which characteristics of the service need improvement.

Both *effectiveness* and *efficiency* are related to the performance in achieving the task goal the service has been built for. Effectiveness is an absolute index which describes to what extent the goal was reached, with respect to the accuracy and completeness of the goals; see, e.g., [14]:

*"Effectiveness:* The accuracy and completeness with which specified users can achieve specified goals in particular environments."

Measures of effectiveness which are reported in literature are, e.g., task success or the kappa metrics [33]. Efficiency, on the other hand, is a relative measure of goal achievement in relation to the resources used [14]:

*"Efficiency:* The resources expended in relation to the accuracy and completeness of goals achieved."

---

<sup>1</sup> ITU-T Rec. G.1000 even defines four different points of view: the customer's requirements for QoS, the service provider's offering of QoS, the QoS achieved or delivered by the provider, and the QoS perceived by the customer.

Commonly used metrics are, e.g., the dialogue duration or the number of turns uttered by the system or by the user.

Efficiency and cognitive demand are criteria characterizing a system with which a user is able to achieve his or her task goals. *Usability*, however, is generally defined in a much broader sense, and describes the capability of the service to be understood, learned and used by specified users under specified conditions. It indicates the suitability of the service to fulfil the user's requirements, includes effectiveness and efficiency of the system, and results in user satisfaction [35]. *User satisfaction* is an indicator of the service's perceived usefulness and usability for the intended user group. It includes whether the user gets the information he/she wants, is comfortable with the service, and gets the information within an acceptable elapsed time [31].

The described notions of quality of a telephone-based spoken dialogue service can be illustrated in terms of a diagram, as has been proposed in [34]; see Figure 2. Apart from the mentioned user factors, four types of factors contribute to the quality perceived by the user: Agent factors (mainly related to the dialogue and the system itself), task factors (related to how the spoken dialogue system captures the task it has been developed for), environmental factors (e.g., factors related to the acoustic environment and to the transmission channel), and contextual factors such as costs, type of access, or the availability. The quality aspects perceived by the user are depicted in the lower part of the diagram.

Environmental, agent, and task factors carry an influence on the *speech input and output quality*, on the *cooperativity* of system behaviour, and on the *symmetry* of the dialogic interaction. Speech input and output quality includes aspects like intelligibility, naturalness, listening-effort required to understand the system messages, or the perceived system understanding. Cooperativity is defined here in the sense of non-violation of principles for cooperative dialogue behaviour, as defined by Grice [20]. It includes the aspects of informativeness, truth and evidence, relevance, manner, background knowledge, and meta-communication handling (i.e., confirmation, clarification, repair and recovery from communication errors); see [6]. The partner asymmetry aspect (differences in interaction behaviour to be attributed to the asymmetry of the interaction partners) is covered by a category called dialogue symmetry. This category also includes the effects of dialogue initiative and interaction control capabilities.

The mentioned quality aspects result in a (more or less) efficient communication (interaction), and in an efficient solution of the task to be carried out. *Communication efficiency* is related to the speed or pace of the interaction, to dialogue conciseness, and to dialogue smoothness. *Task efficiency*, on the other hand, is linked to task success and task ease. Two additional quality aspects are important: the "personality" of the machine agent (politeness, friendliness, naturalness of behaviour) and the effort required from the human user for the interaction (ease of communication, stress/fluster, etc.). These aspects have been subsumed under the term *comfort*.

Communication efficiency, task efficiency and comfort all contribute to the service usability, for which user satisfaction can be seen as an indicator. *Service efficiency*, on the other hand, is influenced by both task efficiency and contextual factors. It is important for the adequacy of the service (for fulfilling the desired task), and for the added value attributed to the service (e.g., in comparison to similar methods for obtaining the same information, like a web interface or a new sticker). Usability, service efficiency, and *economical benefit* result in *utility* of the service, and finally in its *acceptability*.

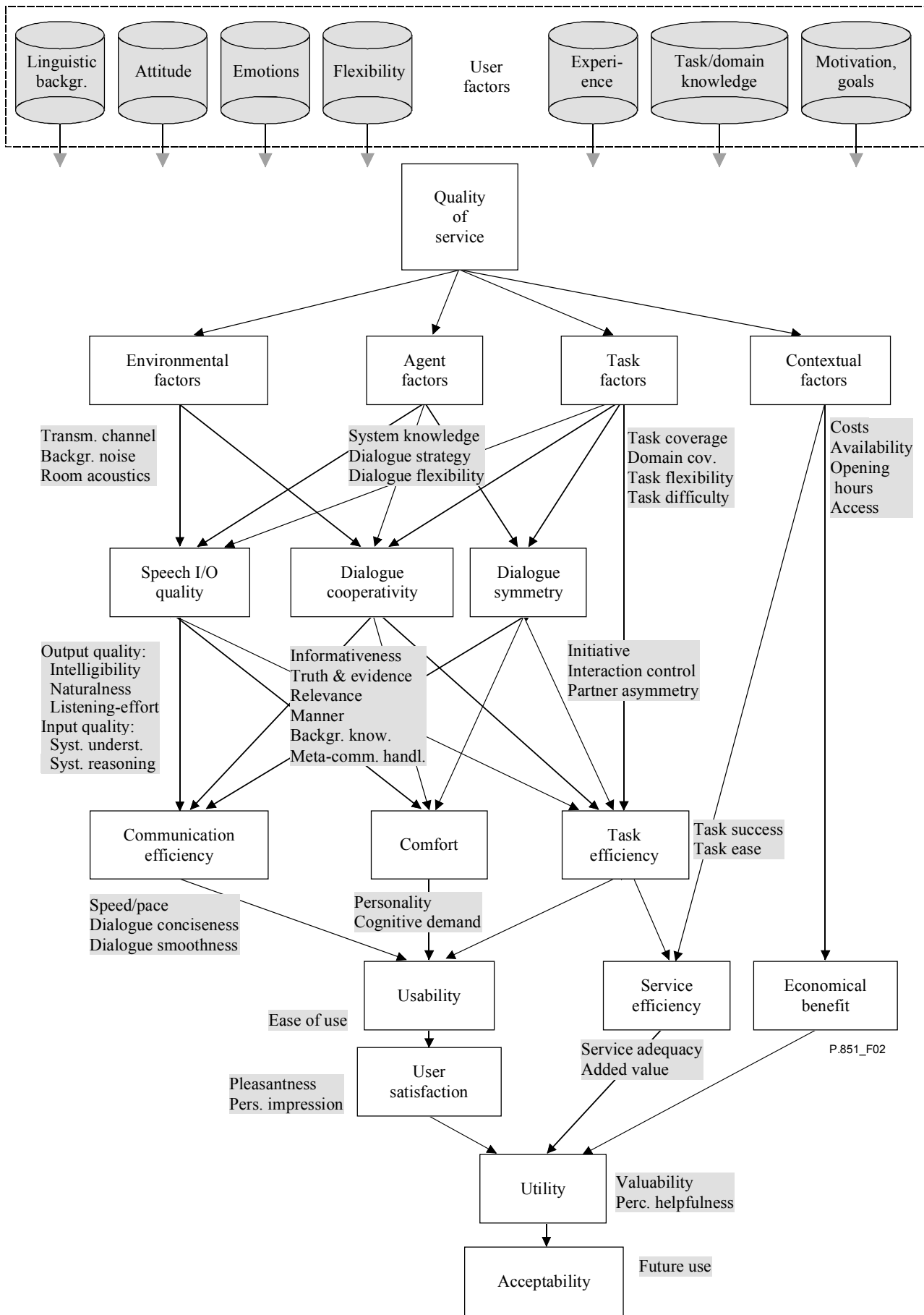


Figure 2/P.851 – Quality aspects and influencing factors; see [34]

#### 4.4 Subjective evaluation methods

Spoken dialogue systems can be assessed on a component level (e.g., with respect to the speech recognizer, to the speech understanding component, or to the speech output component), or with respect to the overall (integrated) system. Analytical assessment on the component level is a valuable source of information in describing how the individual parts of the system fulfil their task. It may, however, sometimes miss the relevant contributors to the overall quality of the service, as perceived by the user. For example, erroneous speech recognition or speech understanding may be compensated for by the discourse processing component, without affecting the overall system quality. For this reason, subjective experiments with real or test users interacting with the entire system are indispensable when the quality of a spoken-dialogue-system-based service is to be determined.

In order to evaluate different aspects of the quality of a spoken-dialogue-system-based service, subjective experiments with human users have to be carried out. These experiments serve two main purposes:

- 1) During the interaction, instrumentally measurable system parameters are collected, and the utterances of the system and of the user are logged. The log-files are submitted to an expert evaluation, the outcome of which is a set of parameters describing specific aspects of the human-machine interaction on the utterance, dialogue and task level, from a system developer's point of view.
- 2) After the interaction, test subjects are given a questionnaire that aims at collecting information about the perceptive quality features which are relevant to form the overall quality impression of the human user. Such experiments can be performed with fully functional systems, or with systems which are still in the development phase and where parts of the system modules have to be simulated. Details on the experimental set-up, the questionnaires, and on usability evaluation methods are given in clauses 6 to 8.

In laboratory experiments, both types of information can be obtained in parallel. In a field test situation with real users, however, instrumentally logged interaction parameters are often the unique source of information for the service provider in order to monitor the quality of the system. The amount of data which can be collected from an operating service may become very large. In this case, it is important to define a core set of metrics which describe system performance, and to have tools at hand which automate a large part of the data analysis process. The task of the human evaluator is then to analyse and interpret this data, and to estimate the effect of the collected performance measures on the quality which would be perceived by a (prototypical) user. Some general considerations about the analysis and interpretation of test results are given in clause 9. Provided that both types of information are available, relationships between interaction parameters and subjective judgements can be established. Such quality prediction models for telephone-based spoken dialogue systems are still under study, and a short discussion is given in clause 9.

As it is common practice for subjective evaluation experiments, the target and the circumstances of an assessment or evaluation experiment should be made explicit, and they should be documented. In the European DISC project, a template has been developed for this purpose [5]. Based on this template and on a classification of methods given in [33], the following criteria can be defined:

- Motivation of assessment/evaluation (e.g., a detailed analysis of the system's recovery mechanisms, or the estimated satisfaction of future users).
- Object of assessment/evaluation (e.g., the speech recognizer, the dialogue manager, or the whole system).
- Environment for assessment/evaluation (e.g., in a controlled laboratory experiment or in a field test).
- Type of measurement methods (e.g., via an instrumental measurement of interaction parameters, or via open or closed quality judgements obtained from the users).

- Symptoms to look for (e.g., user clarification questions or ASR rejections).
- Life cycle phase in which the assessment/evaluation takes place (e.g., for a simulation, a prototype version, or for a fully working system).
- Accessibility of the system and its components (e.g., in a glass box or in a black box approach).
- Reference used for the measurements (e.g., qualitative measures of absolute system performance, or quantitative values with respect to a measurable reference or benchmark).
- Support tools which are available for the assessment/evaluation.

These criteria form a basic set of documentation which should be provided with assessment or evaluation experiments. The documentation may be implemented in terms of an item list as given here, or via a detailed experimental description.

## 5 Spoken dialogue system characterization

Following the schematic of quality aspects given in Figure 2, five types of factors carry an influence on the interaction with a spoken-dialogue-system-based service: *agent factors*, *task factors*, *user factors*, *environmental factors*, and *contextual factors*. These factors will determine the performance of the system (components) and the quality perceived by the user. Thus, they have to be taken into account when conducting and documenting subjective evaluation experiments.

### 5.1 Agent factors

The system as an interaction partner can be characterized in a technical way, namely by defining the characteristics of the individual system components and their interconnection in the sequential structure of Figure 1, or by specifying the agent's operational functions. The most important agent functions to be characterized are the speech recognition capability, the natural language understanding capability, the dialogue management capability, the response generation capability, and the speech output capability. The natural language understanding and the response generation components are closely linked to the neighbouring components, namely the dialogue manager on one side, and the speech recognizer or the speech synthesizer on the other. Thus, the interfaces to these components have to be precisely described.

#### 5.1.1 Speech recognizer characterization

From a functional point of view, speech recognizers can be classified according to the following parameters [41]:

- Vocabulary size, e.g., small, medium, or large vocabulary speech recognizers.
- Vocabulary complexity, e.g., with respect to the confusability of words.
- Speech type, e.g., isolated words, connected words, continuous speech, spontaneous speech including discontinuities such as coughs, hesitations, interruptions, restarts, etc.
- Language: Monolingual or multilingual recognizers, language dependency of recognition results, language portability.
- Speaker dependency, e.g., speaker-dependent, speaker-independent or speaker-adaptive recognizers.
- Type and complexity of grammar. The complexity of a grammar can be determined in terms of its perplexity, which is a measure of how well a word sequence can be predicted by the language model.
- Training method, e.g., multiple training of explicitly uttered isolated words, or embedded training on strings of words of which the starting and ending points are not defined.

On the other hand, speech recognizer components can be described in terms of general technical characteristics which may be implemented differently in individual systems [27]. The following technical characteristics have partly been used in the DISC project:

- Signal capture: Sampling frequency, signal bandwidth, quantization, windowing.
- Feature analysis, e.g., mel-scaled cepstral coefficients, energy, and first or second order derivatives.
- Fundamental speech units, e.g., phone models or word models, modelling of silence or other non-speech sounds.
- Lexicon: Number of entries for each word, with one or several pronunciations; generated either from dictionaries or from grapheme-to-phoneme converters; additional entries for filler words and noises; expected coverage of the vocabulary with respect to the target vocabulary.
- Acoustic model: Type of model, e.g., Multi-Layer-Perceptron (MLP) networks or Hidden Markov Models (HMMs); training data and parameters; post-processing of the model.
- Language model: Type of model, e.g., a statistical N-gram back-off language model, or a context-free grammar; training material, e.g., a large general-purpose training corpus or data collected in a limited experiment; individual word modelling or classes for specific categories (e.g., dates or names); dialogue-state-independent or dialogue-state-dependent models.
- Type of decoder, e.g., HMM-based.
- Degree of use of prosodic information.

### **5.1.2 Language understanding characterization**

The following characteristics are important for the language understanding capability of the system:

- Semantic description of the task, e.g., via slots (attribute-value-pairs).
- Syntactic-semantic analysis: General parsing capability, e.g., full parsing or robust partial parsing; number and complexity of allowed syntax, e.g., the number of alternatives available at a given level.
- Contextual analysis: Number and complexity of rules.
- Interaction with speech recognition and dialogue management modules: Type and amount of input and output information (single hypotheses, ranked lists, etc.), dependency of syntactic-semantic and contextual interpretation on the dialogue state.

### **5.1.3 Dialogue manager characterization**

The approach taken for dialogue management can be defined from a technical point of view, e.g., as a dialogue grammar, a plan-based approach, or a collaborative approach [8], [32]. The most important characteristics of the dialogue manager are the type and amount of knowledge implemented in the manager, the distribution of initiative between the system and the user, and the system's meta-communication (confirmation, clarification, repair and recovery) strategies:

- Dialogue manager knowledge: Dialogue history model (information that has been exchanged in the dialogue so far), task and domain models (scenario, plans, goals and subgoals, objects and their characteristics), world knowledge model, conversational model, and user model.
- Initiative: System-initiative, mixed-initiative, or user-initiative.
- Confirmation strategy: Explicit confirmation, implicit confirmation, "echo" confirmation, summarizing confirmation.
- Repair, clarification and recovery strategies.

- Dialogue manager adaptivity: Constitutive managers that have to learn new notions in their normal operation, or adaptive managers which might include a dynamic user model and might be able to learn the user's communicative strategies.

In addition to the interaction with the user, the interaction with the application system has to be defined, including the interface (programming language), and potential control mechanisms for handling the dynamics of the application system.

#### 5.1.4 Speech generation characterization

Speech generation includes the potential generation of a textual response, and the translation into spoken language. Most systems use one of three types of speech generation: pre-recorded speech, template sentences, or text-to-speech. The following characteristics have to be defined:

- Interaction with the dialogue manager: Type and amount of input information provided by the dialogue manager, e.g., orthographic or annotated text, focus or prosodic information, etc.
- Response generation: Strategy (e.g., formal grammar or simple templates), flexibility (pre-defined vocabulary or open vocabulary), type and amount of information to be included in each utterance, form of the message (syntax, choice of words).
- System voice: Number of voices, gender, professionalism, training, prosodic quality, recording conditions, adaptivity.
- Language: Monolingual or multilingual synthesizers, language identification capability, language portability.
- Type of speech generation: Pre-recorded messages, template sentences, text-to-speech, concept-to-speech.
- Text-to-speech characteristics: Strategy (e.g., model-based or corpus-based), text pre-processing capabilities, model parameters, unit corpus characteristics (types and length of units, coverage of the target vocabulary, etc.), concatenation and/or selection algorithms, prosody generation strategies (fundamental frequency, duration, intensity), etc.
- Contextual characteristics: Speaking style, speaking rate, contextual adaptivity.

#### 5.2 Task factors

The task which is to be carried out by the user is a determining factor of the interaction. It can be characterized with respect to the type of task, task domain, task complexity, task frequency, task consequences, and portability:

- Task type: Can be differentiated according to [6] between:
  - well-structured tasks, having a stereotypical structure that prescribes which piece of information must be exchanged, and often also in which natural order; and
  - ill-structured tasks, containing a large number of optional subtasks whose nature and order are difficult to predict,
 as well as between:
  - homogenous; and
  - heterogeneous, which means *inherently* a combination of several different tasks which are different by their actual nature (e.g., ordering plus information plus device control).
- Task domain: Richness, scalability, number of users that are familiar with the domain, usefulness for the domain, generalizability, etc.
- Task complexity: Number of covered scenarios, maximum number of subgoals, number of subtasks which can be achieved in parallel, minimum number of exchanges necessary to solve the problem, expected complexity of syntax/vocabulary, etc.



- Task frequency, i.e., the frequency with which users can be expected to use the system for the given task. Systems for call routing or flight information (so-called "walk-up-and-use systems") will be used with relatively low frequency, so that potential users cannot be expected to have knowledge about the system, nor to show learning effects (remember behaviour from previous calls) or to accept training.
- Task consequences, e.g., security issues.
- Task portability.

### **5.3 User factors**

In most cases, the characterization of the user is limited to a broad categorization with respect to his/her task, domain and world background, because an exact description of factors important for an individual user (attitude, motivation, emotions, flexibility) cannot be achieved. The following characteristics are often given in evaluation protocols:

- Number of users.
- Age and gender: They are expected to carry an influence on the fundamental frequency and the speech spectrum, but also on the dialogue interaction.
- Level of experience: Novice vs. experienced, occasional user vs. regular user, trained user vs. untrained user.
- Level of expertise in the application domain: Professional users vs. private users.
- Explicit motivation for using the service.
- Physical status, vocal effort, speaking rate, etc.
- Native language, accent, dialect, etc.

For specialized applications, it might be necessary to be more explicit in specifying experience and expertise, e.g., with respect to the knowledge of task goals, the ability to develop strategies to optimize task performance, and the ability to use the devices necessary to perform the task [29].

### **5.4 Environmental factors**

The environment contains the entire physical context of the interaction. A full characterization will generally be impossible, and only the factors which directly affect the speech signal should be described, namely:

- Type and acoustic properties of the user interface.
- Telephone transmission channel: The description can be performed on different levels, e.g., in terms of the transmission, switching and terminal equipment used in the connection, or in terms of the parameters of a reference configuration for network planning; see ITU-T Rec. G.107.
- Room acoustic situation: Includes reverberation, sound coloration, ambient noise levels and spectra, concurrent speakers, etc.

### **5.5 Contextual factors**

These are non-physical factors characterizing the context of use of the service under consideration. Typical factors include:

- Facility of access: Availability of telephone numbers, links to and from other services, etc.
- Service availability: Opening hours, potential restrictions of access.
- Costs: Fixed and time-dependent costs of the interaction, specific account conditions, etc.
- Services with similar functionality: Have to be compared with respect to all other contextual factors.

## 6 Experimental set-up

Subjective interaction experiments with a spoken dialogue system should be set up according to the general rules for conversation-opinion tests which are given in ITU-T Rec. P.800. A more detailed description of the practical issues can be found in the ITU-T Handbook on Telephonometry. This principle applies to the physical conditions of the test cabinets, to the ambient noise characteristics, to the experimental design, and to the general rules for data analysis. In the following sections, only those items are described which are specific to the spoken-dialogue-system interaction, namely the system set-up, the test scenarios, and the test subjects.

Subjective experiments can either be carried out with fully working systems, or with the help of a human experimenter simulating missing parts of the system, or the system as a whole (a so-called "Wizard-of-Oz simulation"). In order to obtain valid and reliable results, the (simulated) system, the test users, and the experimental task have to fulfil several requirements, see clauses 6.1 to 6.3. The interactions are usually logged and annotated by a human expert, so that interaction parameters can be calculated. After each interaction and after the whole test session, questionnaires have to be filled in by the test subjects. These questionnaires allow different aspects of the quality of a spoken-dialogue-system-based service to be quantified. The design of such questionnaires is discussed in clause 7. In clause 8, a short overview of evaluation methods addressing the usability of services is given.

### 6.1 System set-up and Wizard-of-Oz simulation

In order to carry out interaction experiments with human users, a set-up providing the full functionality of the system has to be implemented. The exact nature of the set-up will depend on the availability of system components, and thus on the system development phase. If system components have not yet been implemented, or if an implementation would be unfeasible (e.g., due to the lack of data) or uneconomic, simulation of the respective components or of the system as a whole is required.

The simulation of the interactive system by a human being (the so-called "wizard"), i.e., the Wizard-of-Oz (WoZ) simulation, is a well-accepted technique in the system development phase. At the same time, it serves as a tool for evaluation of the system-in-the-loop, or of the bionic system (half system, half wizard). The idea is to simulate the system taking spoken language as an input, process it in some *principled* way, and generate spoken language responses to the user. In order to provide a realistic telephone service situation, speech input and output should be provided to the users via a simulated or real telephone connection, using a standard user interface. Detailed descriptions of the set-up of WoZ experiments can be found in [16], [6], [3] and [9].

WoZ simulations can be used advantageously in cases where the human capacities are superior to those of computers, as is currently the case for speech understanding or speech output. Because the system can be evaluated before it has been fully set up, the performance of certain system components can be simulated to a degree which is beyond the current state-of-the-art. Thus, an extrapolation to technologies which will be available in the future becomes possible [23]. WoZ simulation allows testing of feasibility, coverage, and adequacy prior to implementation, in a relatively economic way. High degrees of novelty and complex interaction models may be easier to simulate in WoZ than to implement in an implement-test-revise approach. However, the latter is likely to gain ground as standard software and prototyping tools emerge, and in industrial settings where platforms are largely available. WoZ is nevertheless worthwhile if the application is at high risk, and the costs to re-build the system are sufficiently high [6].

The interaction between the human user and the system or the wizard is largely influenced by the five types of factors described in clause 5. From the experimenters' point of view, these factors form variables of the experimental set-up. The variables are either under the control of the experimenter (control variables), accessible and measurable by the experimenter (response variables), or confounding factors where the experimenter has no interest in or no control over. Confounding

factors can be catered for by careful experimental design procedures, namely by a complete or partially complete within-subject design.

A main characteristic of a WoZ simulation is that the test subjects do not realize that the system they are interacting with is simulated. Evidence given in [16] and [9] shows that this goal can be reached in nearly 100% of all cases if the simulation is carefully designed. The most important aspect for the illusion of the subject is the speech input and output capability of the system. Several authors emphasize that the illusion of a dialogue with a computer should be supported by voice distortion, e.g., [17] and [2]. However, other system parameters may be able to cause the same effect, e.g., system directedness.

WoZ simulations should provide a realistic simulation of the system's functionality. Therefore, an exact description of the system functionality and of the system behaviour is needed before the WoZ simulation can be set up. It is important that the wizard adheres to this description, and ignores any superior knowledge and skills which he/she has compared to the system to be tested. This requires a significant amount of training and support for the wizard. Because a human would intuitively use its superior skills, the work of the wizard should be automated as far as possible. A number of tools have been developed for this purpose. They usually consist in a representation of the interaction model, e.g., in terms of a visual graph or of a rapid prototyping software tool, filters for the system input and output channel (e.g., structured audio playback, voice disguise, and recognition simulators), and other support tools like interaction logging (audio, text, video) and domain support (e.g., timetables). Typical examples are described in [23], [15], [9], [6] and [33].

## 6.2 Test scenarios

Because of the lack of a real motivation, laboratory tests often make use of experimental tasks which the subjects have to carry out. The experimental task provides an explicit goal, but this goal should not be confused with a goal which a user would like to reach in a real-life situation. Because of this discrepancy, valid user judgements on system helpfulness and acceptability cannot easily be obtained in a laboratory test set-up.

In a laboratory test, the experimental task is defined by a scenario description. A scenario describes a particular task which the subject has to perform through interaction with the system, e.g., to collect information about a specific train connection, or to search for a specific restaurant [6]. Examples of such scenarios for a restaurant information service are given in Appendix I. Using a pre-defined scenario gives maximum control over the task carried out by the test subjects, while at the same time covering a wide range of possible situations (and possible problems) in the interaction. Scenarios can be intentionally designed to test specific system functionalities (so-called development scenarios), or to cover a wide range of potential interaction situations which is desirable for evaluation. Thus, development scenarios are usually different from evaluation scenarios.

Scenarios help to find different weaknesses in a dialogue, and thereby to increase the usability and acceptability of the final system. They define user goals in terms of the task and the sub-domain addressed in a dialogue, and are a prerequisite to determine whether the user achieved his/her goal. Without a pre-defined scenario it would be extremely difficult to compare results obtained in different dialogues, because the user requests could differ and fall outside the system domain knowledge. If the influence of the task is a factor which has to be investigated in the experiment, the experimenter needs to ensure that all users execute the same tasks. This can only be reached by pre-defined scenarios.

Unfortunately, pre-defined scenarios may have some negative effects on the user's behaviour. Although they do not provide a real-life goal for the test subjects, scenarios prime the users on how to interact with the system. Written scenarios may invite the test subjects to imitate the language given in the scenario, leading to read-aloud instead of spontaneous speech. It has been shown that the choice of scenarios may also influence the solution strategies which are most effective for

resolving the task [43]. Test subjects carrying out pre-defined scenarios are usually not particularly concerned about the response of the system, as they do not really need the information. As a result, task success may not show an important effect on the usability judgements of the test subjects. In addition, it has been reported that test subjects do not always read the instructions carefully, and may ignore or misinterpret key restrictions in the scenarios.

The priming effect on the user's language can be reduced with the help of graphical scenario descriptions; see the examples in Appendix I. A comparison between written and graphical scenarios [6], [13] showed that the massive priming effect of written scenarios can be nearly completely avoided by a graphical representation, but that the diversity of linguistic items (total number of words, number of out-of-vocabulary words) is similar in both cases. Thus, language diversity still has to be assured by collecting utterances from a sufficiently high number of different users, e.g., in a field test situation. Another possibility is to present recorded speech descriptions of the tasks to the test subjects and advise them to take notes [42]. In this way, it is hoped that the involved comprehension and memory processes would leave the subjects with an encoding of the meaning of the task description, but not with a representation of the surface form. An empirical proof of this assumption, however, has not yet been given.

### **6.3 Test subjects**

The general rule for evaluation experiments is that the choice of test subjects should be guided by the purpose of the test. For example, analytic assessment of specific system characteristics will only be possible for trained test subjects who are experts of the system under consideration. However, this group will not be able to judge overall aspects of system quality in a way which would not be influenced by their knowledge of the system. Valid overall quality judgements can only be expected from test subjects which match as close as possible the group of future service users. The general recommendations on the eligibility of test subjects given in ITU-T Rec. P.800 should be respected for subjective interaction experiments with spoken-dialogue-system-based services as well.

An overview of user factors is given in clause 5.3. Some of these factors are responsible for the acoustic and linguistic characteristics of the speech produced by the user, namely age and gender, physical status, speaking rate, vocal effort, native language, dialect, or accent. Because these factors may be very critical for the speech recognition and understanding performance, quality judgements obtained from a user group differing in the acoustic and language characteristics might not reflect the quality which can be expected for the target user group. User groups are however variable and ill-defined. A service which is open to the general public will sooner or later be confronted with a large range of different users. Testing with specified users outside the target user group will therefore provide a measure of system robustness with respect to the user characteristics.

A second group of user factors is related to the experience and expertise with the system, the task, and the domain. Several investigations show that user experience affects a large range of speech and dialogue characteristics. For example, it has been reported that users have the tendency to solve more problems per call when they get used to the system, and that the interaction gets shorter [10]. Other investigations showed that the number of in-vocabulary utterances increased when the users became familiar with the system. At the same time, the task completion rate increased [25]. System familiarity may also lead to a reduced number of user inputs and help messages, and to a reduced transaction time [26], [28].

Users seem to develop specific interaction patterns when they get familiar with a system. It has been postulated that such a pattern is a perceived optimal balance between the effort each individual user has to put into the interaction, and the efficiency with which the interaction takes place [39]. Nearly all users seem to develop stable patterns with the system, but the patterns are not identical for all users. The interaction pattern a user develops may also reflect his or her beliefs of the machine agent, in the sense that the user may have a "cognitive model" of the system which reflects what is regarded as the current system belief [38]. Such a model is partly determined by the

utterances given to the system, and partly by the utterances coming from the system. The user generally assumes that his/her utterances are well understood by the system. In case of misunderstandings, the user gets confused, and dialogue flow problems are likely to occur. Another source of divergence between the user's cognitive model and the system's beliefs is that the system has access to secondary information sources such as an application database. The user may be surprised if confronted with information which he/she did not provide.

## 7 Questionnaires

In order to obtain information about quality features perceived by the user, subjective judgements have to be collected. Two different principles can be applied in the collection: either to identify the relevant quality features in a more or less unguided way, or to quantify pre-determined aspects of quality as responses to closed questions or judgement scaling tasks. Both ways have their advantages and inconveniences: open inquiries help to find quality dimensions which would otherwise remain undetected, and to identify the aspects of quality which are most relevant from the user's point of view. In this way, the interpretation of closed quantitative judgements can be facilitated. Closed questions or scaling tasks facilitate comparison between subjects and experiments, and give an exact means to quantify user perceptions. They can be carried out relatively easily, and untrained subjects often prefer this method of judgement.

Scaling tasks will yield valid and reliable results when two main requirements are satisfied: the items to be judged have to be chosen adequately and meaningfully, and the scaling measurement has to follow well-established rules. Scaling methods are described in detail in the psychometrics literature, e.g., in [21], [12] or [7]. *For rating transmission quality*, the ITU-T recommends absolute category rating (ACR), degradation category rating (DCR) and comparison category rating (CCR) methods; see ITU-T Rec. P.800. *For rating the quality of spoken-dialogue-system-based services*, judgements on continuous rating scales or on different absolute category rating scales are usually solicited from the test subjects. An ACR scale consists of a number of discrete categories one of which has to be chosen by the test subject. The categories are displayed visually and may be labelled with attributes for each category, or for the extreme (left-most and right-most) categories only. Examples for continuous rating scales are depicted in the following clauses. Although the "overall impression" scale is similar to the respective ACR scale for rating transmission quality (see ITU-T Rec. P.800), there is no direct relationship between the ratings, and thus no transformation law linking mean scores obtained on one of the continuous scales to MOS scores used for describing overall transmission quality judgements.

The rating task on both continuous or category scales is often described in terms of a statement (e.g., "The system was easy to understand."), and test subjects have to express their agreement on the statement by marking the respective tick or category of the scale. This method is based on early proposals made by Likert [30], and an exemplary scale is depicted in Figure 3. Numbers are attributed to the categories or to the scale positions, depending on whether the statement is positive (from 1 for "strongly disagree" to 5 for "strongly agree") or negative (from 5 for "strongly disagree" to 1 for "strongly agree"), and the individual ratings are summed up for all subjects. Another possibility is to define self-explaining labels for each category, as it is proposed, e.g., by the ITU-T for speech transmission quality experiments; see ITU-T Rec. P.800.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The system was easy to understand.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

P.851\_F03

**Figure 3/P.851 – Judgement on a statement in a way which was proposed by Likert [30]**

Well-constructed scales will not provide valid information when the quality feature to be judged upon is ill-defined, or when it is not appropriately chosen for the service under consideration. In the following clauses, an exemplary choice of quality aspects is given, each of which can be addressed by a specific question to be rated by the test subjects. Examples of formulated questions or statements are listed as well. The choice of questions to be made for a specific service will depend on the type of service, the tasks which can be carried out, the specific interaction behaviour of the service, the test subject group, as well as on the specific purpose for which the evaluation experiment is carried out. Usually, the number of items to be judged in a single questionnaire should be limited to 15 to 20, so that the test subjects are able to distinguish and reflect the individual items.

In a laboratory set-up, a questionnaire can be given to the test subjects directly after performing an interaction (potentially reflecting the impression after this interaction), and/or after a number of interactions (providing some integration over the past experiences). In a field test, the compilation of the questionnaires cannot be strictly controlled, and the judgement usually refers to a number of interactions carried out in a broadly defined time period. It may occur that negative experiences are more prominent and have a stronger influence in the time integration process than positive ones [11].

### **7.1 Questions related to the user's background**

A number of questions should be answered at the beginning of the test session in order to describe the user and his background which is relevant to the experiment. These questions address the following items:

- Personal information: Age, gender, profession, area of birth, current residence, language proficiency.
- Task-related information: Frequency of task, usual approach when resolving the task (alternative interfaces), motivation, other important task- and domain-related aspects.
- System-related information: Experience with DTMF-based or spoken-dialogue-system-based services, experience with speech technology devices (speech recognition, synthesized speech, etc.).

The following list gives examples of questions which can be asked to the test subjects. They are related to a restaurant information service but can easily be adapted to other tasks and services.

## Questions related to the user's background

### Personal data

Gender:  Female  Male

Age: \_\_\_\_\_ years

Profession/Education: \_\_\_\_\_

Region/City of birth: \_\_\_\_\_

Current residence: \_\_\_\_\_

1 How often do you eat out on an average?

\_\_\_\_ times a week                      \_\_\_\_ times a month                      \_\_\_\_ times a year

2 How would you search for a restaurant when you are in a foreign place (multiple choices possible)?

- |                       |                          |  |                          |
|-----------------------|--------------------------|--|--------------------------|
| 2.1 Magazines         | <input type="checkbox"/> | 2.6 Tips from friends                        | <input type="checkbox"/> |
| 2.2 Commercial flyers | <input type="checkbox"/> | 2.7 Calling an automatic speech-based system | <input type="checkbox"/> |
| 2.3 City guide        | <input type="checkbox"/> | 2.8 Other: _____                             | <input type="checkbox"/> |
| 2.4 Telephone book    | <input type="checkbox"/> |  |                          |
| 2.5 Internet          | <input type="checkbox"/> |  |                          |

3 What is important for you when you decide on a restaurant (multiple choices possible)?

- |                             |                          |                          |                          |
|-----------------------------|--------------------------|--------------------------|--------------------------|
| 3.1 Price                   | <input type="checkbox"/> | 3.6 Ambience             | <input type="checkbox"/> |
| 3.2 Food type               | <input type="checkbox"/> | 3.7 Opening hours        | <input type="checkbox"/> |
| 3.3 Food quality            | <input type="checkbox"/> | 3.8 Service speed        | <input type="checkbox"/> |
| 3.4 Variety of food offered | <input type="checkbox"/> | 3.9 Service friendliness | <input type="checkbox"/> |
| 3.5 Location                | <input type="checkbox"/> | 3.10 Other: _____        | <input type="checkbox"/> |

4 Have you ever used an automatic speech-based information system?

Yes  No

4.1 If yes, on which occasion?

4.1.1 How would you characterize your experience with it?

Extremely bad    Bad    Poor    Fair    Good    Excellent    Ideal

5 Do you have experience with a speech understanding system?

Yes  No

5.1 If yes, what kind of system?

6 Do you have experience with synthesized speech?

Yes  No

6.1 If yes, on which occasion?

7 What information about a restaurant do you want to get from an information system?

## 7.2 Questions related to the individual interaction

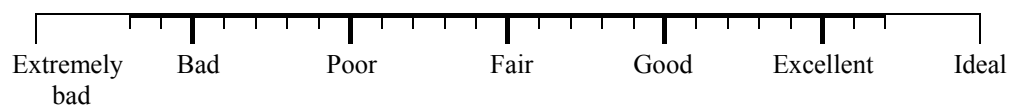
After each interaction with the (simulated) service, the test subjects have to fill in a questionnaire with a number of items related to the individual interaction experience. These items may address the following aspects:

- Information obtained from the system: Availability, accuracy, completeness, consistency, reliability, clarity, and truth of the obtained information, etc.
- Speech input/output capability: Perceived system understanding, frequency of system errors, perceived system reasoning, listening-effort required to understand the system's messages, perceived intelligibility, perceived comprehensibility, etc.
- System's interaction behaviour: Transparency of the interaction, congruence with the user's expectations, flexibility of the interaction, perceived reliability of system processing, distribution of initiative, interaction control capability, confirmation and correction capabilities, recovery from interaction problems, naturalness of the interaction, length of the dialogue, perceived system speed, smoothness of the dialogue, etc.
- Perceived system personality: Friendliness, politeness, etc.
- Impression on the user: Perceived naturalness of the user's own behaviour, pleasantness, cognitive demand put on the user, stress, fluster, etc.
- Perceived task fulfilment: Task success, reliability of task results.

Exemplary questions which address these aspects are given below. The experimenter may select the most appropriate ones for the service under investigation.

### Questions related to the individual interaction

#### Overall impression:



#### Information obtained from the system

- 1 The system provided the desired information.
 

A horizontal scale with 11 tick marks. Labels are: Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree.
- 2 The provided information was ...
 

A horizontal scale with 11 tick marks. Labels are: complete, incomplete.
- 3 The information was ...
 

A horizontal scale with 11 tick marks. Labels are: clear, unclear.
- 4 You would rate the information as ...
 

A horizontal scale with 11 tick marks. Labels are: wrong, true.

#### Communication with the system

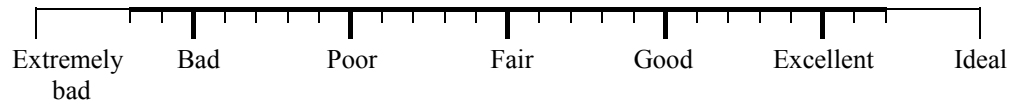
- 5 How well did you feel understood by the system?
 

A horizontal scale with 11 tick marks. Labels are: Extremely bad, Bad, Poor, Fair, Good, Excellent, Ideal.
- 6 You had to concentrate in order to understand what the system expected from you.
 

A horizontal scale with 11 tick marks. Labels are: Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree.

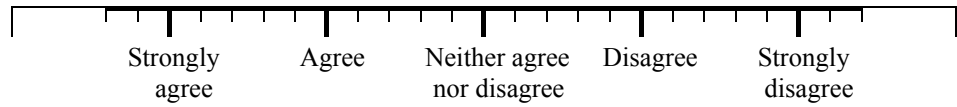


7 How well was the system acoustically intelligible?

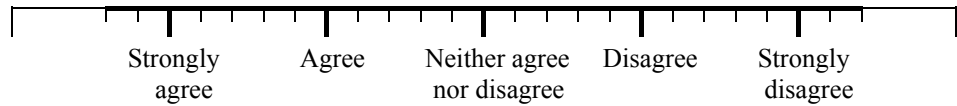


**System behaviour**

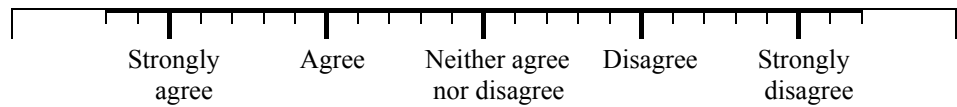
8 You knew at each point of the dialogue what the system expected from you.



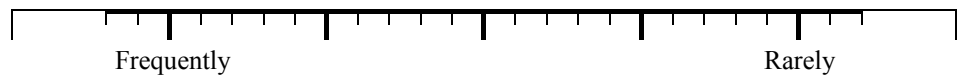
9 In your opinion, the system processed your specifications correctly.



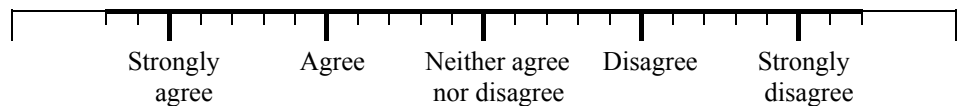
10 The system's behaviour was always as expected.



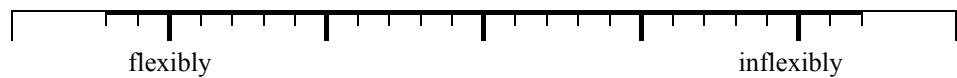
11 How often did the system make mistakes?



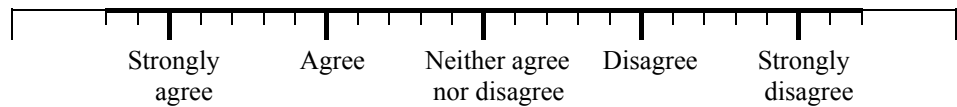
12 The system reacted in the same way as humans do.



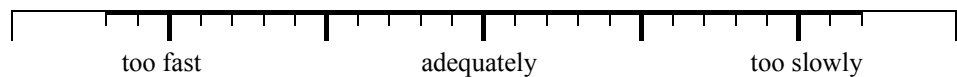
13 The system reacted ...



14 You were able to control the dialogue in the desired way.



15 The system reacted ...



16 The system reacted in a ... way

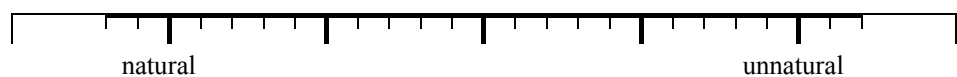


**Dialogue**

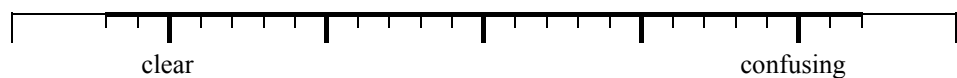
17 The system utterances were ...



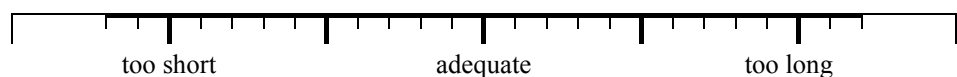
18 You perceived the dialogue as ...



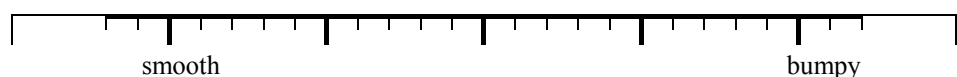
19 The course of the dialogue was ...



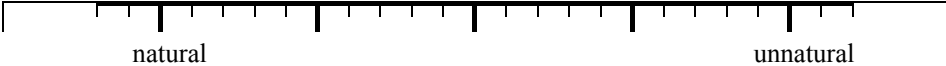
20 The dialogue was ...

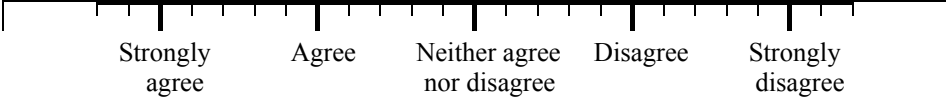


21 The course of the dialogue was ...

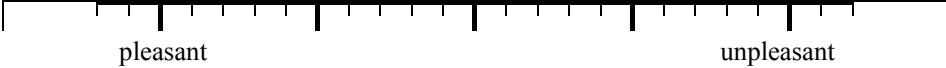


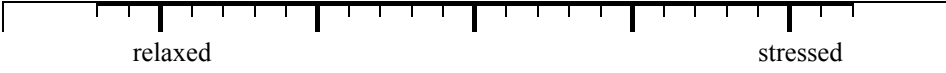
### Your impression of the system

22 The system's voice was ... 

23 Overall, you are satisfied with the dialogue. 

### Personal impression

24 You perceived the dialogue as ... 

25 During the dialogue, you felt ... 

### 7.3 Questions related to the user's overall impression of the system

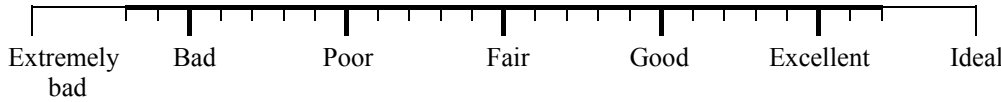
After all of the interactions with the service have been completed, an additional set of questions should be answered by the test subjects, this time referring to their overall experience with the system gained so far. The following items may be included in such a questionnaire:

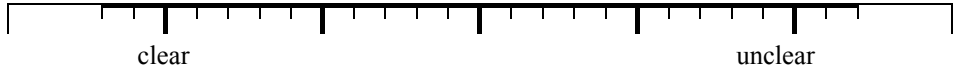
- User's overall impression of the system/service.
- System's manner or expression.
- Perceived system personality: Friendliness, politeness, etc.
- System's correction, recovery and help capabilities.
- Perceived interaction control and initiative.
- Perceived comfort when using the system.
- Perceived task fulfilment: Task success, reliability of task results.
- Perceived usability: Ease of use, ease to learn how to use the system, system habitability.
- User's degree of enjoyment, system likeability.
- Appropriateness and helpfulness of the system for fulfilling the task.
- Added value of the system, in comparison to other interfaces or to a human operator.
- Improvement required before the system may be put into service.
- Expected future use of the service.

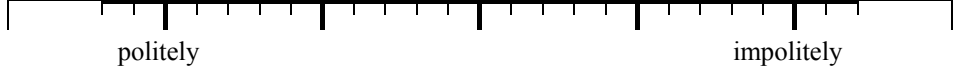
An example of a respective questionnaire for an information service is given below. It may be adapted and extended according to the service under consideration, the task it fulfils, as well as the aims of the experiment.

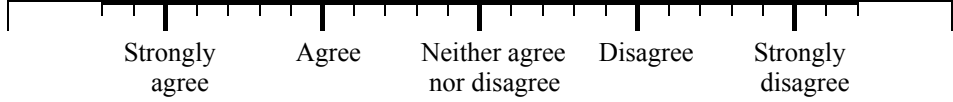
## Questions related to the user's overall impression of the system

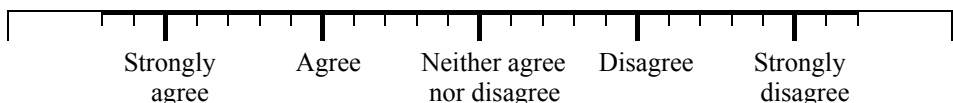
- 1 Overall impression.
 

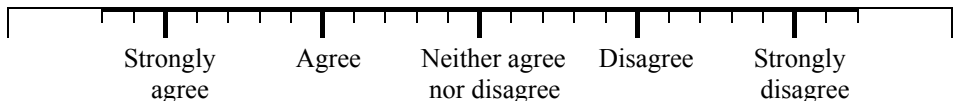

- 2 The system's way of expression was ...
 

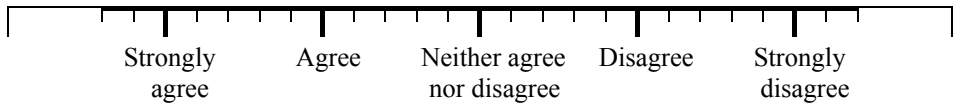

- 3 The system reacted ...
 

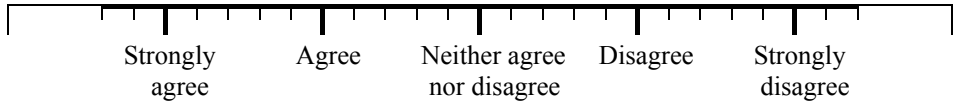

- 4 You would have expected more help from the system.
 



- 5 The system was able to answer all of your questions.
 

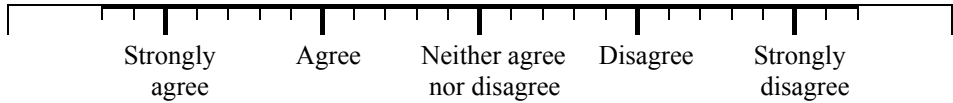

- 6 Misunderstandings could be cleared easily.
 

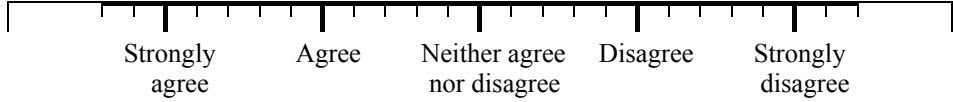

- 7 The system controlled the flow of the dialogue.
 

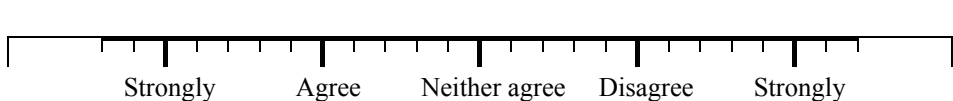

- 8 You were able to handle the system without any problems.
 

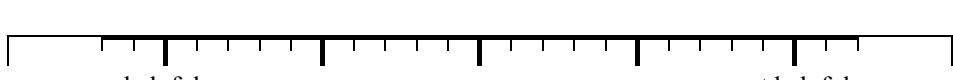

- 9 Regarding the dialogues, you are ...
 



- 10 You enjoyed the dialogues.
 


- 11 You feel adequately informed about the system's possibilities.
 

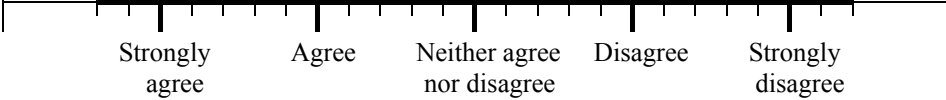

- 12 The telephone calls with the system were worthwhile.
 


- 13 You perceived this possibility for obtaining information as ...
 

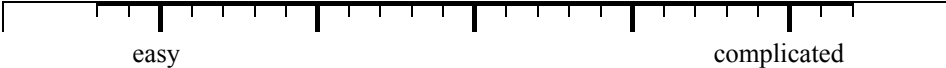

- 14 You rate the system as ....
 



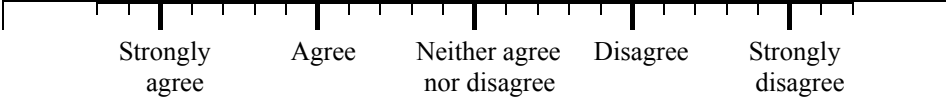
15 You prefer to use another source of information.



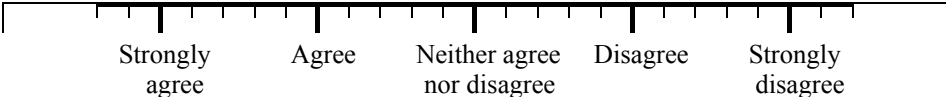
16 The handling of the system was ....



17 You prefer a human operator.



18 In the future, you would use the system again.



19 Which characteristics of the system did you like best?

---

---

20 Which characteristics of the system disturbed you mostly?

---

---

21 Do you have any proposals for system improvement?

---

---

## 8 Usability evaluation

Apart from addressing individual aspects of usability by the described questionnaires, dedicated usability evaluation methods are available. Usability can either be evaluated with real users performing specific tests, or by usability inspection methods with the help of evaluation experts. Both methods are complementary to each other, in that usability inspection methods may be able to detect usability problems which remain overlooked by user testing, and vice versa [36]. In fact, a large degree of non-overlap between the two has been observed. Thus, usability evaluation should combine empirical tests and usability inspections.

Usability inspection methods aim at finding usability problems in an existing user interface design, potentially rating the severity of problems, making recommendations on how to fix the problems, and hereby improving the usability of the system. Such methods allow the knowledge and experience of user interface designers to be easily applied in optimizing new systems. An important part of usability inspection consists of counting and classifying usability problems which are observed in the human-machine interaction. Usability inspection should however not only be efficient in detecting problems, but also in weighting them according to their severity (there is no use in resolving unimportant problems), and especially in suggesting design changes and improvements. Because many inspection methods rely on the design specification rather than on the design implementation, they may be applied relatively early in the system design process.

The following eight types of usability inspection methods may be distinguished [36]:

- *Heuristic evaluation*: This informal method involves usability specialists who judge whether a dialogue element conforms to established usability principles, the so-called heuristics.

- *Guideline reviews*: Inspections where the spoken-dialogue-system-based service is checked for conformance with a comprehensive list of usability guidelines. Because the overall number of guidelines may be very high, this approach requires a high degree of expertise.
- *Pluralistic walkthroughs*: Meetings where users, developers and human factors experts step together through a scenario, discussing usability issues associated with dialogue elements which are involved in each scenario step.
- *Consistency inspections*: An interface is inspected by several designers representing multiple design aspects, and then rated as to whether it is consistent with all design issues.
- *Standards inspections*: An expert investigates a specific interface for compliance with a defined standard.
- *Cognitive walkthroughs*: Simulate a user's problem-solving process at each step in the interaction, and check whether the user's goals and action memory can be assumed to lead to the next correct action. Are typically cast in the form of questions about the relationship between task goals attributed to the user, and the system actions needed to accomplish them.
- *Formal usability inspections*: A formalized method involving a usability inspection team. Each team member has a particular task in the inspection process, e.g., as a moderator, design owner, or inspector. Meetings are organized to prepare and carry out the inspection, and to analyse its results.
- *Feature inspections*: Focuses on the operational functions of the user interface, and whether the provided functions meet the requirements of the intended end users.

Most of these methods are discussed in detail in the respective usability literature [36]. The choice of the right method depends on the objectives of the evaluation, the availability of guidelines, the experience of the evaluator, and time and money constraints.

Usability evaluation with controlled user experiments is the second alternative. Such tests can be carried out either in an "objective, non-intrusive" or in a "subjective, intrusive" way [19]. Non-intrusive methods try to capture the behaviour of the human user in a natural and undisturbed way, e.g., by observation with audiovisual equipment, or by logging with a recording device. Intrusive methods require an active involvement of the users, e.g., by responding to questionnaires or interviews (cf. the last clause), by group discussions, or in a self-descriptive way, i.e., requiring a verbal protocol which reflects the user's thoughts or opinions during or after the interaction. These methods are described in more detail in the usability evaluation literature [14].

## 9 Analysis and interpretation of collected information

The judgements which are obtained on closed rating scales can be analysed by means of barcharts or cumulative distributions. Although the distributions are not necessarily Gaussian, it is common practice to calculate arithmetic mean values (and not medians) over all ratings obtained with a specified system configuration, see ITU-T Rec. P.800 and the ITU-T Handbook on Telephony. For the mean values, confidence limits are evaluated and significance tests performed by conventional analysis of variance (ANOVA). The assumptions underlying an analysis of variance (Gaussian distribution and homogeneity of variances) are not always satisfied; still, this method seems to be robust enough to provide reasonable results also in the case of departures from the statistically ideal conditions. In the case of a statistically significant effect of one of the variates (system configuration and/or voice, test subject, scenario, order of conditions in the experiment, test session, etc.), a post-hoc test can be used to perform pairwise comparisons among the means, and to determine the sources of differences. The Tukey Honestly Significant Difference (HSD) test is recommended for this purpose [40]. When the assumptions underlying a parametric statistics are not satisfied, it is useful to additionally summarize the results in terms of a median or mode, and to use non-parametric tests like the one according to Kruskal and Wallis for comparison.



#### Scenario No. 4

You plan to eat out in XXX. Because your favourite restaurant is closed for holidays, ask the system for a restaurant.

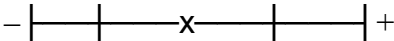
Please write down first which specifications you want to give to the system.

If the system is unable to find a matching restaurant, please search for an alternative until the system indicates at least one restaurant.

Restaurant name(s): \_\_\_\_\_

#### Scenario No. 5

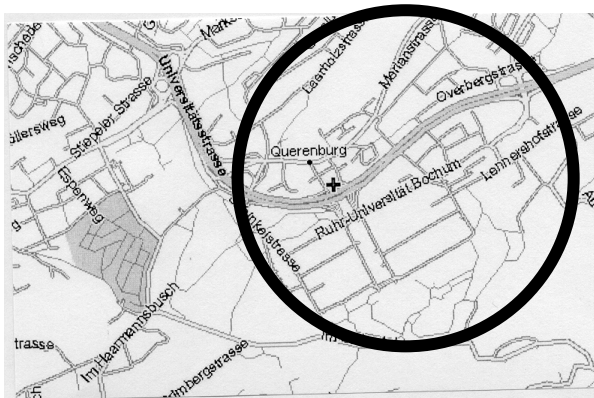
Please gather your information from the following hints:

Price: 

Type of food:



Location:



Restaurant name(s): \_\_\_\_\_

## BIBLIOGRAPHY

- [1] ALLEN (J.), FERGUSON (G.), STENT (A.): An Architecture for More Realistic Conversational Systems, *Proc. of Intelligent User Interfaces 2001 (IUI-01)*, 1-8 Santa Fe NM (2001).
- [2] AMALBERTI (R.), CARBONELL (N.), FALZON (P.): User Representations of Computer Systems in Human-Computer Speech Interaction, *Int. Journal on Man-Machine Studies*, 38, 547-566 (1993).
- [3] ANDERNACH (T.), DEVILLE (G.), MORTIER (L.): The Design of a Real World Wizard of Oz Experiment for a Speech Driven Telephone Directory Information Service, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'93)*, 2, 1165-1168, Berlin (1993).
- [4] ANTONIOL (G.), FIUTEM (R.), LAZZARI (G.), DE MORI, (R.): System Architectures and Applications, *Spoken Dialogues with Computers*, R. de Mori, ed., 583-609, Academic Press, London (1998).
- [5] BERNSEN (N.O.), DYBKJÆR (L.): A Methodology for Evaluating Spoken Dialogue Systems and Their Components, *Proc. 2nd Int. Conf. on Language Resources and Evaluation (LREC 2000)*, 2, 183-188, Athens (2000).
- [6] BERNSEN (N.O.), DYBKJÆR (H.), DYBKJÆR (L.): Designing Interactive Speech Systems: From First Ideas to User Testing, *Springer*, Berlin (1998).
- [7] BORG (I.), STAUFENBIEL (T.): Theorien und Methoden der Skalierung: Eine Einführung, *Verlag Hans Huber*, Bern (1993).
- [8] CHURCHER (G.E.), ATWELL (E.S.), SOUTER (C.): Dialogue Management Systems: A Survey and Overview, *Report 97.06, School of Computer Studies, University of Leeds*, Leeds (1997).
- [9] DAHLBÄCK (N.), JÖNSSON (A.), AHRENBERG (L.): Wizard of Oz Studies – Why and How? *Knowledge-Based Systems*, 6(4), 258-266 (1993).
- [10] DELOGU (C.), DI CARLO (A.), SEMENTINA (C.), STECCONI (S.): A Methodology for Evaluating Human-Machine Spoken Language Interaction, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'93)*, 2, 1427-1430, Berlin (1993).
- [11] DUNCANSON (J.P.): The Average Telephone Call Is Better Than the Average Telephone Call, *The Public Opinion Quarterly*, 33(1), 112-116 (1969).
- [12] DUNN-RANKIN (P.): Scaling Methods, *Lawrence Erlbaum Assoc.*, Hillsdale NJ (1983).
- [13] DYBKJÆR (L.), BERNSEN (N.O.), DYBKJÆR (H.): Scenario Design for Spoken Language Dialogue Systems Development, *Proc. ESCA Workshop on Spoken Dialogue Systems*, P. Dalsgaard, L.B. Larsen, L. Boves and I. Thomsen, eds., 93-96, Vigsø (1995).
- [14] ETSI Technical Report ETR 095: Human Factors (HF); Guide for Usability Evaluations of Telecommunication Systems and Services, *European Telecommunications Standards Institute*, Sophia Antipolis (1993).
- [15] FOSTER (J.C.), DUTTON (R.), JACK (M.A.), LOVE (S.), NAIRN (I.A.), VERGEYNST (N.), STENTIFORD (F.W.M.): Intelligent Dialogues in Automated Telephone Services, *Interactive Speech Technology: Human Factor Issues in the Application of Speech Input/Output to Computers*, C. Baber and J.M. Noyes, eds., 167-175, Taylor and Francis, London (1993).
- [16] FRASER (N.M.), GILBERT (G.N.): Simulating Speech Systems, *Computer Speech and Language*, 5, 81-99 (1991).



- [17] FRASER (N.M.), GILBERT (G.N.): Effects of System Voice Quality on User Utterances in Speech Dialogue Systems, *Proc. 2nd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'91)*, 1, 57-60, Genova (1991).
- [18] GIBBON (D.), MOORE (R.), WINSKY (R.), eds.: Handbook on Standards and Resources for Spoken Language Systems, *Mouton de Gruyter*, Berlin (1997).
- [19] GLEISS (N.): Usability – Concepts and Evaluation, *TELE (English Edition)*, 2/92, 24-30, Swedish Telecommunications Administration, Stockholm (1992).
- [20] GRICE (H.P.): Logic and Conversation, *Syntax and Semantics, Vol. 3: Speech Acts*, P. Cole and J.L. Morgan, eds., 41-58, Academic Press, New York NY (1975).
- [21] GUILFORD (J.P.): Psychometric Methods, *McGraw-Hill Book Company*, New York NY (1954).
- [22] HONE (K.S.), GRAHAM (R.): Towards a Tool for Subjective Assessment of Speech System Interfaces (SASSI), *Natural Language Engineering*, 6(3-4), 287-303 (2000).
- [23] JACK (M.A.), FOSTER (J.C.), STENTIFORD (F.W.M.): Intelligent Dialogues in Automated Telephone Services, *Proc. 2nd Int. Conf. on Spoken Language Processing (ICSLP'91)*, 1, 715-718, Banff (1992).
- [24] JEKOSCH (U.): *Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*, Habilitation thesis (unpublished), University/GH Essen, Essen (2000).
- [25] KAMM (C.), NARAYANAN (S.), DUTTON (D.), RITENOUR (R.): Evaluating Spoken Dialogue Systems for Telecommunication Services, *Proc. 5th Europ. Conf. on Speech Communication and Technology (EUROSPEECH'97)*, 4, 2203-2206, Rhodes (1997).
- [26] LAMEL (L.), BENNACEF (S.), GAUVAIN (J.L.), DARTIGUES (H.), TEMEM (J.N.): User Evaluation of the MASK Kiosk, *Speech Communication*, 38, 131-139 (2002).
- [27] LAMEL (L.), MINKER (W.), PAROUBEK (P.): Towards Best Practice in the Development and Evaluation of Speech Recognition Components for a Spoken Language Dialogue System, *Natural Language Engineering*, 6(3-4), 305-322 (2000).
- [28] LAMEL (L.), BENNACEF (S.), GAUVAIN (J.L.), DARTIGUES (H.), TEMEM (J.N.): User Evaluation of the MASK Kiosk, *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, 7, 2875-2878, Sydney (1998).
- [29] LIFE (M.A.), LEE (B.P.), LONG (J.B.): Assessing the Usability of Future Speech Technology: Towards a Method, *Proc. of SPEECH'88*, 7th FASE Symposium, 4, 1297-1304, Edinburgh (1988).
- [30] LIKERT (R.): A Technique for the Measurement of Attitudes, *Archives of Psychology*, 140, 1-55 (1932).
- [31] MAIER (E.), MAST (A.), LUPERFOY (S.): Overview. Dialogue Processing in Spoken Language Systems, *Proc. of the ECAI'96 Workshop*, Budapest, E. Maier, M. Mast and S. LuperFoy, eds., Lecture Notes in Artificial Intelligence No. 1236, 1-13, Springer, Berlin (1997).
- [32] McTEAR (M.F.): Spoken Dialogue Technology: Enabling the Conversational Interface, *ACM Computing Surveys*, 34(1), 90-169 (2002).
- [33] MÖLLER (S.): Quality of Telephone-Based Spoken Dialogue Systems, *Habilitation thesis, Institute of Communication Acoustics*, Ruhr-University, Bochum (to appear) (2003).

- [34] MÖLLER (S.): A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems, *Proc. 3rd SIGdial Workshop on Discourse and Dialogue*, 142-153, Philadelphia PA (2002).
- [35] MÖLLER (S.): Assessment and Prediction of Speech Quality in Telecommunications, *Kluwer Academic Publ.*, Boston MA (2000).
- [36] NIELSEN (J.), MACK (R.L.), eds.: Usability Inspection Methods, *John Wiley & Sons*, New York NY (1994).
- [37] SENEFF (S.): Galaxy-II: A Reference Architecture for Conversational System Development, *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, 3, 931-934, Sydney (1998).
- [38] SOUVIGNIER (B.), KELLNER (A.), RUEBER (B.), SCHRAMM (H.), SEIDE (F.): The Thoughtful Elephant: Strategies for Spoken Dialog Systems, *IEEE Trans. Speech and Audio Processing*, 8(1), 51-62 (2000).
- [39] STURM (J.), BAKX (I.), CRANEN (B.), TERKEN (J.), WANG (F.): The Effect of Prolonged Use of Multimodal Interaction, *Proc. ISCA Workshop on Multi-Modal Dialogue in Mobile Environments*, L. Dybkjær, E. André, W. Minker and P. Heisterkamp, eds., 1-15, Kloster Irsee (2002).
- [40] TUKEY (J.W.): Exploratory Data Analysis, *Addison-Wesley*, Reading MA (1997).
- [41] VAN LEEUWEN (D.), STEENEKEN (H.): Assessment of Recognition Systems, *Handbook on Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore and R. Winsky, eds., 381-407, Mouton de Gruyter, Berlin (1997).
- [42] WALKER (M.A.), RUDNICKY (A.), PRASAD (R.), ABERDEEN (J.), BRATT (E.O.), GAROFOLO (J.), HASTIE (H.), LE (A.), PELLOM (B.), POTAMIANOS (A.), PASSONNEAU (R.), ROUKOS (S.), SANDERS (G.), SENEFF (S.), STALLARD (D.): DARPA Communicator: Cross System Results for the 2001 Evaluation, *Proc. 7th Int. Conf. on Spoken Language Processing (ICSLP 2002)*, 1, 269-272, Denver CO (2002).
- [43] WALKER (M.A.), FROMER (J.), DI FABBRIZIO (G.), MESTEL (C.), HINDLE (D.): What Can I Say? Evaluating a Spoken Language Interface to Email, *Human Factors in Computing Systems. CHI'98 Conf. Proc.*, Los Angeles CA, 582-589, Assoc. for Computing Machinery (ACM), New York NY (1998).
- [44] WALKER (M.A.), LITMAN (D.J.), KAMM (C.A.), ABELLA (A.): PARADISE: A Framework for Evaluating Spoken Dialogue Agents, *Proc. of the ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics*, 271-280 (1997).
- [45] ZUE (V.), SENEFF (S.), GLASS (J.R.), POLIFRONI (J.), PAO (C.), HAZEN (T.J.), HETHERINGTON (L.): JUPITER: A Telephone-Based Conversational Interface to Weather Information, *IEEE Trans. Speech and Audio Processing*, 8(1), 85-96 (2000).







## SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series B	Means of expression: definitions, symbols, classification
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	TMN and network maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
<b>Series P</b>	<b>Telephone transmission quality, telephone installations, local line networks</b>
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communications
Series Y	Global information infrastructure, Internet protocol aspects and Next Generation Networks
Series Z	Languages and general software aspects for telecommunication systems