



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
Methods for objective and subjective assessment of quality	Series	P.80 P.800
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

Page

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

Copyright International Telecommunication Union/ITU Telecommunication Sector

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.

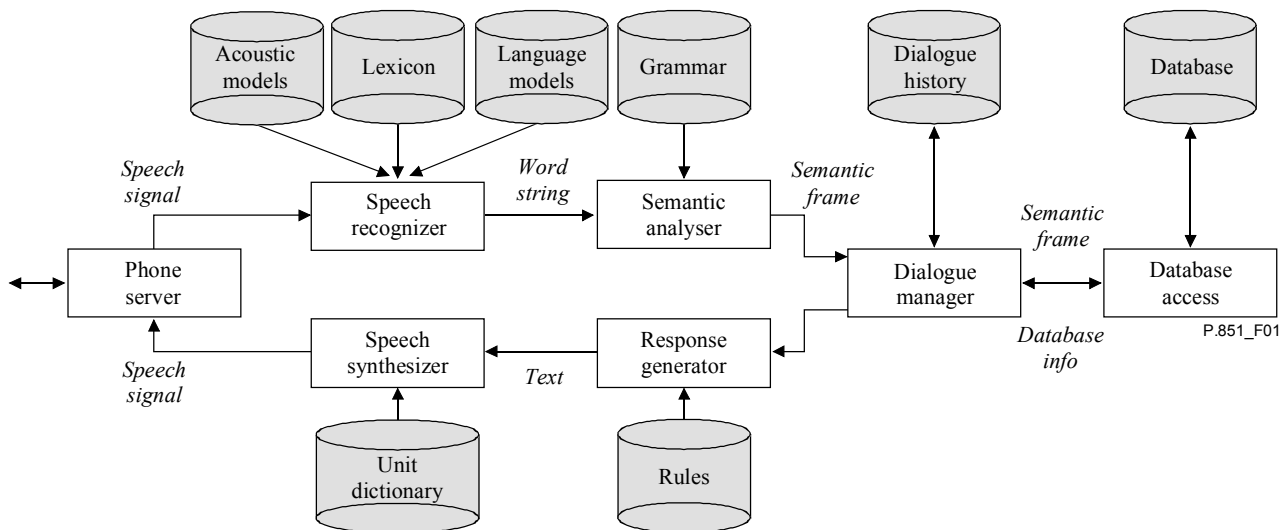


Figure 1/P.851 – Sequential structure of a telephone-based spoken dialogue system [27], [33]

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 user¹. 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."

¹ 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*.

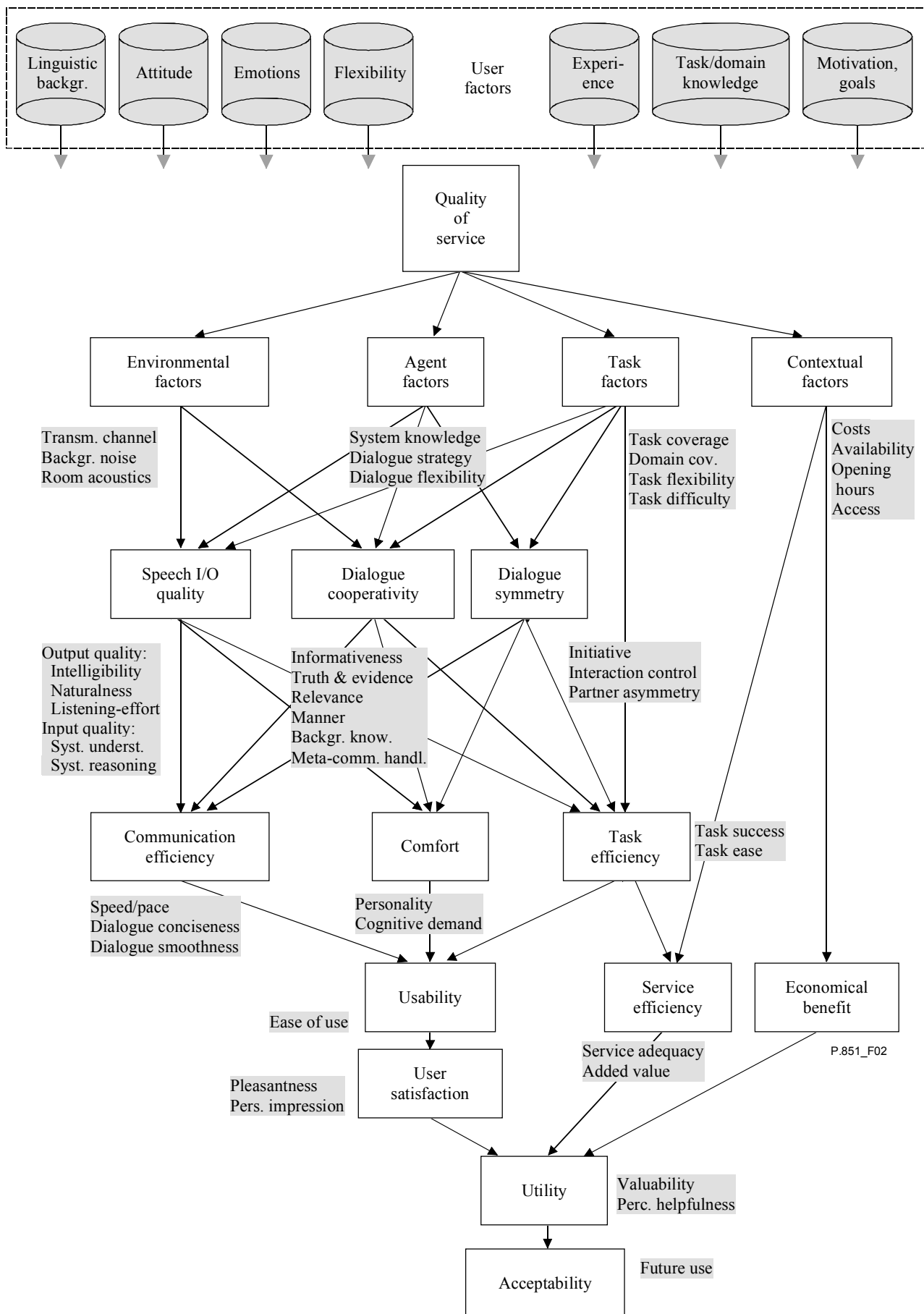


Figure 2/P.851 – Quality aspects and influencing factors; see [34]