International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Series P
Supplement 25
(01/2011)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Parameters describing the interaction with multimodal dialogue systems

ITU-T P-series Recommendations – Supplement 25

ITU-T P-SERIES RECOMMENDATIONS

TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	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 speech quality	Series	P.80
		P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100

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

Supplement 25 to ITU-T P-series Recommendations

Parameters describing the interaction with multimodal dialogue systems

Summary

Supplement 25 to the ITU-T P-series Recommendations provides definitions for a set of parameters which can be extracted from services which rely on multimodal dialogue systems. The parameters can be extracted from logged (test) user interactions with the service under consideration. They quantify the flow of the interaction, the behaviour of the user and the system, and the performance of the speech technology devices involved in the interaction. They provide useful information for system development, optimization and maintenance, and are complementary to subjective quality judgments. The list is an amendment and extension of the respective list of parameters for speech-based services which is given in Supplement 24 to the ITU-T P-series Recommendations.

History

Edition	Recommendation	Approval	Study Group
1.0	ITU-T P Suppl. 25	2011-01-27	12

Keywords

Assessment, automatic speech recognition, automatic speech understanding, dialogue management, gesture recognition, interaction parameter, multimodal dialogue system, speech generation.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). 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 publication, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this publication is voluntary. However, the publication may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the publication 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 publication is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this publication 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 publication development process.

As of the date of approval of this publication, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this publication. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Supplement 25 to ITU-T P-series Recommendations

Parameters describing the interaction with multimodal dialogue systems

1 Scope

This supplement describes parameters providing information on the interaction with services which are based on multimodal dialogue systems, as seen by the system developer and service operator. Multimodal dialogue systems addressed by this supplement enable a multimodal interaction with a human user. Such systems offer one or more modalities for input (e.g., speech, gesture, touch) and one or more output modalities (e.g., a graphical user interface, spoken output, an embodied conversational agent) and may have automatic speech, gesture or touch recognition, speech understanding, a fusion module, dialogue management, response generation, a fission module and speech, graphical or audiovisual output capabilities. They may provide access to information stored in a database, or allow different types of transactions to be performed, and they are frequently offered on smart-phone platforms.

The parameters defined here quantify the flow of the interaction, the behaviour of the user and the system, and the performance of the devices involved in the interaction. For extracting all parameters, the multimodal dialogue system has to be accessible as a glass box; still, some parameters may also be extracted in a black-box approach, i.e., without access to the individual system components. The extraction can partially be performed automatically, and partially relies on a human expert transcribing and annotating interaction log files. The parameters address system performance from a system developer's point-of-view; thus, they provide complementary information to subjective evaluation experiments. Further guidance on subjective evaluation methods in general and on the assessment of speech output devices and spoken dialogue systems is available in [ITU-T P.800], [ITU-T P.85] and [ITU-T P.851], and in the Handbook on Telephonometry. This guidance, however, does not yet cover multimodal systems.

2 References

[ITU-T P.10]	Recommendation ITU-T P.10/G.100 (2006), *Vocabulary for performance and quality of service*.
[ITU-T P.85]	Recommendation ITU-T P.85 (1994), *A method for subjective performance assessment of the quality of speech voice output devices*.
[ITU-T P.800]	Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.
[ITU-T P.851]	Recommendation ITU-T P.851 (2003), *Subjective quality evaluation of telephone services based on spoken dialogue systems*.
[ITU-T Handbook]	ITU-T Handbook on Telephonometry (1993).
[IEC 60268-16]	IEC Standard 60268-16 (1998), *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index*.
[Bernsen-1]	Bernsen, N.O., Dybkjær, H., Dybkjær, L. (1998), *Designing interactive speech systems: From first ideas to user testing*. Springer, DE-Berlin.
[Bernsen-2]	Bernsen, N.O. (2002), *Multimodality in Language and Speech Systems – From Theory to Design Support Tool*. In Granström, B., House, D., and Karlsson, I. (Eds.): Multimodality in Language and Speech Systems, Dordrecht, Kluwer Academic Publishers, 93-148.

[Billi] Billi, R., Castagneri, G., Danieli, M. (1996), *Field trial evaluations of two different information inquiry systems*. In: Proc. 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'96), US-Basking Ridge NJ, 129-134.

[Boros] Boros, M., Eckert, W., Gallwitz, F., Gorz, G., Hanrieder, G., Niemann, H. (1996), *Towards understanding spontaneous speech: Word accuracy vs. concept accuracy*. In: Proc. 4th Int. Conf. on Spoken Language Processing (ICSLP'96), IEEE, US-Piscataway NJ, 2, 1009-1012.

[Carletta] Carletta, J. (1996), *Assessing agreement of classification tasks: The kappa statistics*. Computational Linguistics, 22(2), 249-254.

[Chu] Chu, M., Peng, H. (2001), *An objective measure for estimating MOS of synthesized speech*. In: Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavia), DK-Aalborg, 3, 2087-2090.

[Cookson] Cookson, S. (1988), *Final evaluation of VODIS – Voice operated data inquiry system*. In: Proc. of Speech'88, 7th FASE Symposium, UK-Edinburgh, 4, 1311-1320.

[Danieli] Danieli, M., Gerbino, E. (1995), *Metrics for evaluating dialogue strategies in a spoken language system*. In: Empirical Methods in Discourse Interpretation and Generation. Papers from the 1995 AAAI Symposium, US-Stanford CA, AAAI Press, US-Menlo Park CA, 34-39.

[Fraser] Fraser, N. (1997), *Assessment of interactive systems*. In: Handbook on Standards and Resources for Spoken Language Systems (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, DE-Berlin, 564-615.

[Gerbino] Gerbino, E., Baggia, P., Ciaramella, A., Rullent, C. (1993), *Test and evaluation of a spoken dialogue system*. In: Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP'93), IEEE, US-Piscataway NJ, 2, 135-138.

[Gibbon] Gibbon, D., Moore, R., Winski, R., Eds. (2000), *Handbook on Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, DE-Berlin.

[Glass] Glass, J., Polifroni, J., Seneff, S., Zue, V. (2000), *Data collection and performance evaluation of spoken dialogue systems: The MIT experience*. In: Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000), CN-Beijing, 4, 1-4.

[Goodine] Goodine, D., Hirschman, L., Polifroni, J., Seneff, S., Zue, V. (1992), *Evaluating interactive spoken language systems*. In: Proc. 2nd Int. Conf. on Spoken Language Processing (ICSLP'92), CA-Banff, 1, 201-204.

[Grice] Grice, H.P. (1975), *Logic and conversation*. In: Syntax and Semantics, Vol. 3: Speech Acts (P. Cole and J.L. Morgan, eds.), Academic Press, US-New York NY, 41-58.

[Hirschman] Hirschman, L., Pao, C. (1993), *The cost of errors in a spoken language system*. In: Proc. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93), DE-Berlin, 2, 1419-1422.

[Kamm] Kamm, C.A., Litman, D.J., Walker, M.A. (1998), *From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems*. In: Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98), AU-Sydney, 4, 1211-1214.

[Kühnel] Kühnel, C., Weiss, B., Möller, S., (2010), *Parameters describing multimodal interaction – Definitions and three usage scenarios.* Procedures of the 11th Annual Conference of the ISCA (Interspeech 2010), JP – Tokyo 2014-2017.

[Möller] Möller, S. (2005), *Quality of telephone-based spoken dialogue systems*. Springer, US-New York NY.

[Nigay] Nigay L. and Coutaz J. (1993), *A design space for multimodal systems: concurrent processing and data fusion*. In: Proc. Of the INTERACT and CHI, 172-178.

[NIST SRST] NIST Speech Recognition Scoring Toolkit (2001), *Speech recognition scoring toolkit*. National Institute of Standards and technology, http://www.nist.gov/speech/tools, US-Gaithersburg MD.

[Perakakis] Perakakis, M and Potamianos, A. (2008), *Multimodal system evaluation using modality efficiency and synergy metrics*. In: *Proc. of the 10th international conference on Multimodal interfaces* (ICMI '08). ACM, New York, NY, USA, 9-16.

[Picone-1] Picone, J., Doddington, G.R., Pallett, D.S. (1990), *Phone-mediated word alignment for speech recognition evaluation*. IEEE Trans. Acoustics Speech and Signal Processing, 38(3), 559-562.

[Picone-2] Picone, J., Goudie-Marshall, K.M., Doddington, G.R., Fisher, W. (1986), *Automatic text alignment for speech system evaluation*. IEEE Trans. Acoustics Speech and Signal Processing, 34(4), 780-784.

[Polifroni] Polifroni, J., Hirschman, L., Seneff, S., Zue, V. (1992), *Experiments in evaluating interactive spoken language systems*. In: Proc. DARPA Speech and Natural Language Workshop, US-Harriman CA, 28-33.

[Price] Price, P.J., Hirschman, L., Shriberg, E., Wade, E. (1992), *Subject-based evaluation measures for interactive spoken language systems*. In: Proc. DARPA Speech and Natural Language Workshop, US-Harriman CA, 34-39.

[San-Segundo] San-Segundo, R., Montero, J.M., Colás, J., Gutiérrez, J., Ramos, J.M., Pardo, J.M. (2001), *Methodology for dialogue design in telephone-based spoken dialogue systems: A Spanish train information system*. In: Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavia), DK-Aalborg, 3, 2165-2168.

[Simpson] Simpson, A., Fraser, N.M. (1993), *Black box and glass box evaluation of the SUNDIAL system*. In: Proc. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93), DE-Berlin, 2, 1423-1426.

[Skowronek] Skowronek, J. (2002), *Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstequalität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem*. Diploma thesis (unpublished), Institut für Kommunikationsakustik, Ruhr-Universität, DE-Bochum.

[Strik-1] Strik, H., Cucchiarini, C., Kessens, J.M. (2001), *Comparing the performance of two CSRs: How to determine the significance level of the differences*. In: Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavia), DK-Aalborg, 3, 2091-2094.

[Strik-2] Strik, H., Cucchiarini, C., Kessens, J.M. (2000), *Comparing the recognition performance of CSRs: In search of an adequate metric and statistical significance test*. In: Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000), CN-Beijing, 4, 740-743.

[van Leeuwen] van Leeuwen, D., Steeneken, H. (1997), *Assessment of recognition systems*. In: Handbook on Standards and Resources for Spoken Language Systems (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, DE-Berlin, 381-407.

[Walker-1] Walker, M.A., Fromer, J., di Fabbrizio, G., Mestel, C., Hindle, D. (1998), *What can I say?: Evaluating a spoken language interface to email*. In: Human Factors in Computing Systems. CHI'98 Conference Proc., US-Los Angeles CA, ACM, US-New York NY, 582-589.

[Walker-2] Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A. (1998), *Evaluating spoken dialogue agents with PARADISE: Two case studies*. Computer Speech and Language, 12(3), 317-347.

[Walker-3] Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A. (1997), *PARADISE: A framework for evaluating spoken dialogue agents*. In: Proc. of the 35th Ann. Meeting of the Assoc. for Computational Linguistics, ES-Madrid, 271-280.

[Zue] Zue, V., Seneff, S., Glass, J.R., Polifroni, J., Pao, C., Hazen, T.J., Hetherington, L. (2000), *JUPITER: A telephone-based conversational interface for weather information*. IEEE Trans. Speech and Audio Processing, 8(1), 85-96.

3 Definitions

For definitions not listed here, please refer to [ITU-T P.10].

3.1 **barge-in**: The ability of a human to speak over a system prompt or system output [Gibbon].

3.2 **cooperativity**: The term cooperativity describes the system's ability to engage in a cooperative interaction with the human user, following the principles for cooperative behaviour by Grice [Grice] and their adaptation to spoken dialogue systems by Bernsen et al. [Bernsen-1]. It describes the extent to which the system output is informative, truthful, relevant, well mannered, fitting to the user's background knowledge, and the extent to which the system is able to handle meta-communication [Möller].

3.3 **dialogue**: A conversation or an exchange of information. As an evaluation unit: One of several possible paths through the dialogue structure.

3.4 **efficiency**: Measures of the accuracy and completeness of system tasks relative to the resources (e.g., time, human effort) used to achieve the specific system tasks.

3.5 **element**: Smallest information carrying bit of a turn: A word in case of spoken interaction, a click on a GUI, an information changed on a GUI, etc.

3.6 **exchange**: A pair of contiguous and related turns, one spoken by each party in the dialogue [Fraser].

3.7 **functionality**: Capability of the system to provide functions which meet stated and implied needs when the system is used under specific conditions.

3.8 **gesture**: Non-verbal communication in which visible three-dimensional bodily actions or two-dimensional surface gestures on a touch screen communicate particular messages. A gesture is every kind of user input executed with body parts (e.g., hand, arm, or head) that is neither handwriting nor keyboard input.

3.9 **meta-communication**: The communication about communication, e.g., for resolving misunderstandings ("Did I understand you right?") or for reaching agreement on the use of the language.

3.10 performance: The ability of a unit to provide the function it has been designed for.

3.11 speech technology: The discipline concerned with the research and development of spoken language input and output systems using contributions from the neighbouring disciplines of acoustics, electrical engineering, statistics, phonetics, natural language processing, and involving system requirements specification, design, implementation and evaluation, corpus and linguistic resource processing, and consumer oriented product evaluation [Gibbon].

3.12 spoken dialogue system: A computer system with which human users interact via spoken language on a turn-by-turn basis.

3.13 task: All the activities which a user must develop in order to attain a fixed objective in some domain.

3.14 task-oriented dialogue: A dialogue concerning a specific subject, aiming at an explicit goal (such as resolving a problem or obtaining specific information) [Fraser].

3.15 transaction: The part of a dialogue devoted to a single high-level task (e.g., making a travel booking or checking a bank account balance). A transaction may be coextensive with a dialogue or a dialogue may consist of more than one transaction [Fraser].

3.16 touch: Input via a touch screen, usually a button press, to be distinguished from two-dimensional surface gestures.

3.17 turn: Input respectively output by the user respectively the system, from when the input/output begins until the end or until the other party takes over. Equals an utterance in case of spoken dialogue systems.

3.18 utterance: A stretch of speech, spoken by one party in a dialogue, from when this party starts speaking until another party definitely takes over [Bernsen-1].

4 Abbreviations

This supplement uses the following abbreviations:

ASR	Automatic Speech Recognition
AVM	Attribute-Value Matrix
AVP	Attribute-Value Pair
DARPA	Defense Advanced Research Projects Agency
DP	Dynamic Programming
DTMF	Dual Tone Multiple Frequency
IVR	Interactive Voice Response
MDS	Multimodal Dialogue System
MOS	Mean Opinion Score
SDS	Spoken Dialogue System
WoZ	Wizard-of-Oz

5 Conventions

None.

6 Introduction

Multimodal dialogue systems (MDSs), i.e., computer systems with which human users interact via different modalities such as spoken language, gestures or touch on a turn-by-turn basis, may be part of modern telephone networks. They enable access to databases and transactions, e.g., for obtaining train or airline timetable information, stock exchange rates, tourist information, or to perform bank account operations or make hotel reservations, or enable the user to control different devices or services remotely. Frequently, such systems are offered on smart-phone platforms. In contrast to interactive voice response (IVR) systems with DTMF input and spoken dialogue systems (SDS), MDSs offer interaction in multiple (complementary or redundant) modalities. These include spoken language as in SDS, graphical interfaces to be operated with 2-dimensional gestures (touch, stylus, mouse, etc.), or 3-dimensional gestures (e.g., motion of a device, movements of the hands, face, or whole body) recognized for example via cameras or motion sensors.

In order to evaluate the quality of services which rely on SDSs from a user's perspective, ITU-T published Recommendation P.851 in 2003. That Recommendation describes methods for conducting subjective evaluation experiments in order to determine quality from a user's point-of-view, taking the SDS as a black box. With the help of experiments carried out according to [ITU-T P.851], valuable information on quality, as perceived by the user, may be obtained. However, it may be difficult to determine how the individual system components contribute to the overall quality experienced by the user, e.g., to determine which component needs improvement in case of interaction problems. Thus, the evaluation should be complemented with information that addresses the system performance from a system designer's and service operator's point-of-view.

System-related information may be described in terms of so-called interaction parameters. Such parameters help to quantify the flow of the interaction, the behaviour of the user and the system, and the performance of the speech technology devices involved in the interaction. They address system performance from a system developer's and service operator's point-of-view, and thus provide complementary information to subjective evaluation data. For extracting some of the parameters, the spoken dialogue system has to be accessible as a glass box. Other parameters may also be extracted in a black-box approach, i.e., without an access to the individual system components.

This supplement provides a collection of interaction parameters which have been used for evaluating SDSs in the past 15 years and MDSs in the last eight years. The listed parameters are related to the overall communication of information between user and system, the meta-communication in case of misunderstandings, the cooperativity of the system, the task which can be carried out with the help of the system, and the system's speech input capabilities. No parametric description is yet available for speech output quality (e.g., with respect to synthesized speech quality). The collection of SDS parameters is based on the theoretical work described in [Möller]. The additional parameters applicable for MDS have been published in [Kühnel].

Not all of the interaction parameters proved to be in a direct relationship to the perceived quality of MDS-based services. In fact, correlations between individual parameters and users' quality judgments are generally quite moderate. Still, it will be advantageous to dispose of a large set of parameters describing the interaction between user and system – in this way capturing most of the information which is potentially relevant for perceived quality from a system designer's perspective. Such parameters provide useful information for system development, optimization, and maintenance.

If the parameters are applied in evaluation experiments at different test sites, it may become possible to estimate the relationship between parameters and perceived quality for a wide range of systems and services. In this way, it may become possible to develop algorithms for predicting quality on the basis of interaction parameters. Work in this direction is still under way within ITU-T and elsewhere.

7 Characteristics of interaction parameters

Interaction parameters can be extracted when real or test users interact with the service. The extraction can be performed partly instrumentally and partly with the help of log files which have to be transcribed and annotated by a human expert. Simple parameters, like the duration of the interaction or of single turns, can usually be measured fully instrumentally with appropriate algorithms. On the other hand, human transcription and annotation is necessary when not only the surface form (speech signals) is addressed, but also the contents and meaning of system or user utterances (e.g., to determine a word or concept accuracy).

MDSs are of such high complexity that a description of system behaviour and a comparison between systems or system versions needs to be based on a multitude of different parameters [Simpson]. As a consequence, both (instrumental and expert-based) ways of collecting interaction parameters should be followed in order to get as much information as possible. Based on the collected information, spoken dialogue services can be optimized and maintained very efficiently.

Because interaction parameters are based on data which has been collected in an interaction between user and system, they are influenced by the characteristics of the system, of the user, and of the interaction between both. These influences can usually not be separated, because the user's behaviour is strongly influenced by the behaviour of the system (e.g., the questions asked by the system), and vice versa (e.g., the vocabulary and speaking style of the user influences the system's recognition and understanding accuracy). Consequently, interaction parameters strongly reflect the characteristics of the user group they have been collected with.

Interaction parameters are either determined in a laboratory test setting under controlled conditions or in a field test. In the latter case, it may not be possible to extract all parameters, because not all necessary information can be gathered. For example, if the success of a task-oriented interaction (e.g., collection of a train timetable) is to be determined, then it is necessary to know about the exact aims of the user. Such information can only be collected in a laboratory setting, e.g., in the way it is described in [ITU-T P.851]. In case that the fully integrated system is not yet available, it is possible to collect parameters from a so-called "Wizard-of-Oz" (WoZ) simulation, where a human experimenter replaces missing parts of the system under test. The characteristics of such a simulation have to be taken into account when interpreting the obtained parameters.

Interaction parameters can be calculated on a word level, on a sentence or utterance level, or on the level of a full interaction or dialogue. In case of word or utterance level parameters, average values are often calculated for each dialogue. The parameters collected with a specific group of users may be analysed with respect to the impact of the system (version), the user group, and the experimental setting (scenarios, test environment, etc.), using standard statistical methods. A characterization of these influences can be found in [ITU-T P.851].

8 Review of interaction parameters

Based on a broad literature survey, parameters were identified which have been used in different assessment and evaluation experiments during the past 15 years. The respective literature can be found in [Billi] [Boros] [Carletta] [Cookson] [Danieli] [Fraser] [Gerbino] [Glass] [Goodine] [Hirschman] [Kamm] [Polifroni] [Price] [San-Segundo] [Simpson] [Skowronek] [Strik-1] [Strik-2] [van Leeuwen] [Walker-2] [Nigay] [Zue], and the parameters have been summarized in [Möller]. The parameters can broadly be classified as follows:

- dialogue- and communication-related parameters;
- meta-communication-related parameters;
- cooperativity-related parameters;
- task-related parameters;

- input-related parameters;
- output-related parameters.

These categories will be briefly discussed in the following clauses. For each category, the respective parameters will be listed, together with a definition, the interaction level addressed by the parameter (word, utterance or dialogue), as well as the measurement method (instrumental or based on expert annotation).

8.1 Dialogue- and communication-related parameters

Parameters which refer to the overall dialogue and to the communication of information give a very rough indication of how the interaction takes place. They do not specify the communicative function of each individual turn in detail. Parameters belonging to this category are listed in Table 1, and include duration-related parameters (overall dialogue duration, duration of system and user turns, system and user response delay), and element- and turn-related parameters (average number of system and user turns, average number of elements per system and per user turn, number of system and user questions).

Two parameters which have been proposed in [Glass] are worth noting: The query density gives an indication of how efficiently a user can provide new information to a system, and the concept efficiency describes how efficiently the system can absorb this information from the user. These parameters also refer to the system's language understanding capability, but they have been included in this clause because they result from the system's interaction capabilities as a whole, and not purely from the language understanding capabilities.

The parameters relative modality efficiency and multimodal synergy proposed by Perakakis and Potamianos [Perakakis] should help identify suboptimal use of modalities due to poor interface design or information asymmetries and measure the quality of modality fusion.

All parameters in this category are of global character and refer to the dialogue as a whole, although they are partly calculated on an utterance level. Global parameters are sometimes problematic, because the individual differences in cognitive skill may be large in relation to the system-originated differences, and because subjects might learn strategies for task solution which have a significant impact on global parameters.

Table 1 – Dialogue- and communication-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. meth.
DD	dialogue duration	Overall duration of a dialogue in [ms], see e.g., [Fraser] [Cookson] [Goodine] [Polifroni]	Dial.	Instr.
STD	system turn duration	Average duration of a system turn, from the beginning of system output to the end of system output, in [ms]. In the case of speech-only output a turn is an utterance, i.e., a stretch of speech spoken by one party in the dialogue [Fraser].	Turn	Instr.
UTD	user turn duration	Average duration of a user turn, from the beginning of observable user input to the end of user input, in [ms] [Fraser]. In the case of GUI interaction from the beginning of the movement towards the GUI to the end of the click, in the case of a gesture from the beginning to the end of the movement.	Turn	Instr.

Table 1 – Dialogue- and communication-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. meth.
SRD	system response delay	Average delay of a system response, from the end of user input to the beginning of system output, in [ms] [Price], e.g., from button press to display of new GUI.	Turn	Instr.
SFD	system feedback delay	Average delay of system feedback, from the end of user input to the beginning of system feedback, e.g., from button press to display of loading status in terms of clock, etc.	Turn	Instr.
URD	user response delay	Average delay of a user response, from the end of system output to the beginning of user input, in [ms] [Price], e.g., from the display of GUI.	Turn	Instr.
# turns	number of turns	Overall number of turns in a dialogue. [Walker-2] Feedback does not count as a turn; only in the case where missing or negative feedback interrupts the user input and/or leads to a repetition, it is counted as system turn and as an error. Can be annotated per modality.	Dial.	Instr./ expert.
# system turns	number of system turns	Overall number of system turns in a dialogue [Walker-2], can be annotated per modality	Dial.	Instr./ expert.
# user turns	number of user turns	Overall number of user turns in a dialogue [Walker-2], can be annotated per modality	Dial.	Instr./ expert.
EPST	elements per system turn	Average number of elements per system turn in a dialogue [Cookson], i.e., words, or sounds, or number of information carrying bits changed in a GUI.	Turn	Instr./ expert.
EPUT	elements per user turn	Average number of elements per user turn in a dialogue [Cookson], i.e., words, gestures, keys pressed.	Turn	Instr./ expert.
# system questions	number of system questions	Overall number of questions from the system per dialogue, display of list is implicit question to Select.	Dial.	Expert.
# user questions	number of user questions	Overall number of questions from the user per dialogue [Goodine] [Polifroni].	Dial.	Expert.
QD	query density	Average number of new concepts (slots, see clause 8.4) introduced per user query. Being n_d the number of dialogues, $n_q(i)$ the total number of user queries in the i-th dialogue, and $n_u(i)$ the number of unique concepts correctly "understood" by the system in the i-th dialogue, then $$QD = \frac{1}{n_d} \sum_{i=1}^{n_d} \frac{n_u(i)}{n_q(i)}$$ A concept is not counted to $n_u(i)$ if the system already understood it in one of the previous utterances [Glass].	Set of dial.	Expert.

Table 1 – Dialogue- and communication-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. meth.
CE	concept efficiency	Average number of turns which are necessary for each concept to be "understood" by the system. Being n_d the number of dialogues, $n_u(i)$ the number of unique concepts correctly "understood" by the system in the i-th dialogue, and $n_c(i)$ the total number of concepts in the i-th dialogue, then $$CE = \frac{1}{n_d} \sum_{i=1}^{N_d} \frac{n_u(i)}{n_c(i)}$$ A concept is counted whenever it was uttered by the user and was not already understood by the system [Glass].	Set of dial.	Expert.
# SMC	number of system output modality changes	Overall number of modality changes by the system.	Dial.	Instr.
# UMC	number of user input modality changes	Overall number of modality changes by the user.	Dial.	Instr.
RME	relative modality efficiency	Number of information bits that are communicated correctly via modality mod per time unit or per turn [Perakakis]: $$RME_{\text{mod}} = \frac{\frac{N_{\text{mod}}}{T_{\text{mod}}}}{\sum \frac{N_i}{T_i}}$$ N_{mod}: Number of information bits communicated with modality mod. T_{mod}: Overall time or number of turns spent using modality mod.	Dial.	Instr. /expert
MS	multimodal synergy	Percent improvement in terms of time-to-task-completion achieved by the multimodal system compared to a system randomly combining modalities or compared to the average time-to-completion of a corresponding unimodal system [Perakakis].	Dial.	Instr. /expert

8.2 Meta-communication-related parameters

Meta-communication, i.e., the communication about communication, is particularly important for the interaction with systems which have limited recognition, understanding, and reasoning capabilities. In this case, correction and clarification turns or even sub-dialogues are needed to recover from misunderstandings.

The parameters belonging to this group quantify the number of system and user turns which are part of meta-communication. Most of the parameters are calculated as the absolute number of turns in a dialogue which relate to a specific interaction problem and are then averaged over a set of dialogues. They include the number of help requests from the user, of time-out prompts from the system, of user utterances or gestures rejected by the system in the case that no semantic content could be extracted (ASR and GR rejections), of diagnostic system error messages, of barge-in attempts from the user, and of user attempts to cancel a previous action.

The ability of the system (and of the user) to recover from interaction problems can be described in two ways: Either explicitly by the correction rate, i.e., the percentage of all (system or user) turns which are primarily concerned with rectifying an interaction problem, or implicitly with the implicit recovery parameter, which quantifies the capacity of the system to regain utterances which have partially failed to be recognized or understood.

In contrast to the global measures, most meta-communication-related parameters describe the function of system and user utterances in the communication process. Thus, most parameters have to be determined with the help of an annotating expert. The parameters are listed in Table 2.

Table 2 – Meta-communication-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. meth.
# *help request*	number of help requests from the user	Overall number of user help requests in a dialogue. A user help request is labelled by the annotation expert if the user explicitly asks for help or accessed the system help via a GUI or gesture, etc. This request may be formulated as a question (e.g., "What are the available options?") or as a statement (e.g., "Give me the available options!") [Walker-2].	Turn	Expert.
# *system help*	number of diagnostic system help messages	Overall number of help messages generated by the system in a dialogue. A help message can be a system utterance which informs the user about available options at a certain point in the dialogue or can be displayed via a GUI.	Turn	Instr./ expert.
# *time-out*	number of time-out prompts	Overall number of time-out prompts, due to no response from the user, in a dialogue [Walker-2].	Turn	Instr.
# *ASR rejection*	number of ASR rejections	Overall number of ASR rejections in a dialogue. An ASR rejection is defined as a system prompt indicating that the system was unable to "hear" or to "understand" the user, i.e., that the system was unable to extract any meaning from a user utterance [Walker-2].	Turn	Instr.
# *GR rejection*	number of gesture recognition rejection	Overall number of GR rejections in a dialogue. Defined by a system prompt or feedback indicating that the system was unable to "understand" the users' gesture.	Turn	Instr.

Table 2 – Meta-communication-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. meth.
# system error	number of diagnostic system error messages	Overall number of diagnostic error messages from the system in a dialogue. A diagnostic error message is defined as a system utterance, feedback or graphical output in which the system indicates that it is unable to perform a certain task or to provide a certain information [Price].	Turn	Instr./ expert.
# barge-in	number of user barge-in attempts	Overall number of user barge-in attempts in a dialogue. A user barge-in attempt is counted when the user intentionally addresses the system while the system is still speaking. In this definition, user utterances which are not intended to influence the course of the dialogue (laughing, expressions of anger or politeness) are not counted as barge-ins [Walker-2].	Turn	Expert.
# cancel	number of user cancel attempts	Overall number of user cancel attempts in a dialogue. A user turn is classified as a cancel attempt if the user tries to restart the dialogue from the beginning, or if he/she explicitly wants to step one or several levels backwards in the dialogue hierarchy [Kamm] [San-Segundo].	Turn	Expert.
SCT, SCR	number of system correction turns, system correction rate	Overall number (*SCT*) or percentage (*SCR*) of all system turns in a dialogue which are primarily concerned with rectifying a "trouble", thus not contributing new propositional content and interrupting the dialogue flow. A "trouble" may be caused by speech recognition or understanding errors, or by illogical, contradictory, or undefined user utterances. In case that the user does not give an answer to a system question, the corresponding system answer is labelled as a system correction turn, except when the user asks for an information or action which is not supported by the current system functionality [Fraser] [Simpson] [Gerbino] [Danieli].	Turn	Expert.
UCT, UCR	number of user correction turns, user correction rate	Overall number (*UCT*) or percentage (*UCR*) of all user turns in a dialogue which are primarily concerned with rectifying a "trouble", thus not contributing new propositional content and interrupting the dialogue flow (see *SCT*, *SCR*) [Fraser] [Simpson] [Gerbino] [Danieli].	Turn	Expert.
IR	implicit recovery	Capacity of the system to recover from user utterances for which the speech recognition or understanding process partly failed. Determined by labelling the partially parsed utterances (see definition of *PA*:*PA* in clause 8.5) as to whether the system response was "appropriate" or not: $$IR = \frac{\#\ utterances\ with\ appropriate\ system\ answer}{PA : PA}$$ For the definition of "appropriateness" see clause 8.3 [Danieli].	Turn	Expert.

8.3 Cooperativity-related parameters

Cooperativity has been identified as a key aspect for a successful interaction with a spoken dialogue system [Bernsen-1]. Unfortunately, it is difficult to quantify whether a system behaves cooperatively or not. Several of the dialogue- and meta-communication-related parameters somehow relate to system cooperativity, but they do not attempt to quantify this aspect.

Direct measures of cooperativity are the contextual appropriateness parameters introduced by Simpson and Fraser [Simpson]. Each system utterance has to be judged by a number of experts as to whether it violates one or more of Grice's maxims for cooperativity, see [Grice]:

- Quantity of information: Make your contribution as informative as required (for the current purpose of the exchange); do not make your contribution more informative than is required.
- Quality: Try to make your contribution one that is true; do not say what you believe to be false; do not say that for which you lack adequate evidence.
- Relation: Be relevant.
- Manner: Be perspicuous; avoid obscurity of expression; avoid ambiguity; be brief (avoid unnecessary prolixity); be orderly.

These principles have been stated more precisely by Bernsen and Dybkjær [Bernsen-1] with respect to spoken dialogue systems.

The utterances are classified into the categories of appropriate (not violating Grice's maxims), inappropriate (violating one or more maxims), appropriate/inappropriate (the experts cannot reach agreement in their classification), incomprehensible (the content of the utterance cannot be discerned in the dialogue context), or total failure (no linguistic response from the system). It has to be noted that the classification is not always straightforward, and that interpretation principles may be necessary.

Table 3 – Cooperativity-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. meth.
CA:AP, *CA:IA*, *CA:TF*, *CA:IC*, *%CA:AP*, *%CA:IA*, *%CA:TF*, *%CA:IC*	contextual appropriateness	Overall number or percentage of system utterances which are judged to be appropriate in their immediate dialogue context. Determined by labelling utterances according to whether they violate one or more of Grice's maxims for cooperativity: • *CA:AP*: Appropriate, not violating Grice's maxims, not unexpectedly conspicuous or marked in some way. • *CA:IA*: Inappropriate, violating one or more of Grice's maxims. • *CA:TF*: Total failure, no linguistic response. • *CA:IC*: Incomprehensible, content cannot be discerned by the annotation expert. For more details see [Simpson] [Fraser] [Gerbino]; the classification is similar to the one adopted in [Hirschman].	Turn	Expert.

8.4 Task-related parameters

Current state-of-the-art services enable task-orientated interactions between system and user, and task success is a key issue for the usefulness of a service. Task success may best be determined in a laboratory situation where explicit tasks are given to the test subjects, see [ITU-T P.851]. However, realistic measures of task success have to take into account potential deviations from the scenario by the user, either because he/she did not pay attention to the instructions given in the scenario, because of his/her inattentiveness to the system utterances, or because the task was irresolvable and had to be modified in the course of the dialogue.

Modification of the experimental task is considered in most definitions of task success which are reported in the literature. Success may be reached by simply providing the right answer to the constraints set in the instructions, by constraint relaxation from the system or from the user (or both), or by spotting that no solution exists for the defined task. Task failure may be tentatively attributed to the system's or to the user's behaviour, the latter however being influenced by the behaviour of the system.

A different approach to determine task success is the κ coefficient. It assumes a speech-understanding approach which is based on attributes (concepts, slots) for which allowed values have to be assigned in the course of the dialogue between system and user. The pairs of attributes and assigned values are called attribute-value pairs (AVPs). A set of all available attributes together with the values assigned by the task (a so-called attribute-value matrix (AVM)) completely describes a task which can be carried out with the help of the system. In order to determine the κ coefficient, a confusion matrix $M(i,j)$ is set up for the attributes in the key (scenario definition) and in the reported solution (log file of the dialogue). Then, the agreement between key and solution $P(A)$ and the chance or likelihood of agreement $P(E)$ can be calculated from this matrix, see Table 4. $M(i,j)$ can be calculated for individual dialogues, or for a set of dialogues which belong to a specific system or system configuration.

The κ coefficient relies on the availability of a simple task coding scheme, namely in terms of an AVM. However, some tasks cannot be characterized as easily. In that case, more elaborated approaches to task success are needed, approaches which usually depend on the type of task under consideration.

Table 4 – Task-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. Meth.
TS	task success	Label of task success according to whether the user has reached his/her goal by the end of a dialogue, provided that this goal could be reached with the help of the system. The labels indicate whether the goal was reached or not, and the assumed source of problems: • *TS:S*: Succeeded (task for which solutions exist) • *TS:SCs*: Succeeded with constraint relaxation by the system • *TS:SCu*: Succeeded with constraint relaxation by the user • *TS:SCsCu*: Succeeded with constraint relaxation both from the system and from the user • *TS:SN*: Succeeded in spotting that no solution exists • *TS:Fs*: Failed because of the system's behaviour, due to system adequacies • *TS:Fu*: Failed because of the user's behaviour, due to non-cooperative user behaviour See also [Fraser] [Danieli] [Simpson].	Dial.	Expert.
κ	kappa coefficient	Percentage of task completion according to the kappa statistics. Determined on the basis of the correctness of the result AVM reached at the end of a dialogue with respect to the scenario (key) AVM. A confusion matrix $M(i,j)$ is set up for the attributes in the result and in the key, with T the number of counts in M, and t_i the sum of counts in column I of M. Then $$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$ where $P(A)$ is the proportion of times that the AVM of the actual dialogue and the key agree, $P(A) = \sum_{i=1}^{n} \frac{M(i,i)}{T}$. $P(E)$ can be estimated from the proportion of times that they are expected to agree by chance, $$P(E) = \sum_{i=1}^{n} (\frac{t_i}{T})^2$$ See [Nigay] [Carletta].	Dial. or set of dial.	Expert.

8.5 Input-related parameters

Input-modality appropriateness

Multimodal dialogue systems may offer a set of modalities for user input. These modalities may be used sequentially, simultaneously, or compositely [Nigay]. Depending on the content, the environment and the user can be determined (for example, guided by modality properties as described in [Bernsen-2]) if the offered input modalities are appropriate for every given turn. This can be annotated per modality or, in the case of composite input, for the multimodal input as a whole. In the first case, each modality can be appropriate or inappropriate. In the second case, the multimodal input can be appropriate, partially appropriate, or inappropriate.

Table 5 – Input-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. Meth.
IMA:AP, IMA:PA, IMA:IA, %IMA:AP, %IMA:PA, %IMA:IA	input modality appropriateness	Overall number or percentage of chosen input modalities which are judged to be appropriate in their immediate dialogue context. Determined by labelling user input according to whether they violate one or more modality properties [Bernsen-2]: • *IMA:AP*: Appropriate. • *IMA:PA*: Partially appropriate. • *IMA:IA*: Inappropriate.	Turn	Expert.

Speech, handwriting, and keyboard input

The speech input capability of a dialogue system is determined by its capability to recognize words and utterances, and to extract the meaning from the recognized string (so-called "speech understanding"). For automatic speech recognition, two approaches have to be distinguished: Word recognizers are able to extract single words from the user's speech when spoken in isolation (isolated word recognition) or continuously (keyword spotting). On the other hand, continuous speech recognizers are able to recognize whole sentences or utterances. A similar distinction can be made for handwriting recognition on the basis of words or symbols. Speech understanding is often performed on the basis of attribute-value pairs, see the previous clause. The parameters described in the following paragraph address both speech recognition and speech understanding.

Continuous speech recognizers generally provide a word string hypothesis as an output. In order to judge whether the string correctly represents what has been said, a reference transcription has to be provided by the transcribing expert. For each utterance, hypothesized and reference string are first aligned on a word level, using a dynamic programming (DP) matching algorithm [Picone-1] [Picone-2]. On the basis of the alignment, the number of correctly determined words c_w, of substitutions s_w, of insertions i_w, and of deletions d_w is counted. These counts can be related to the total number of words in the reference n_w, resulting in two alternative measures of recognition performance, the word error rate *WER* and the word accuracy *WA*, see Table 6.

Complementary performance measures can be defined on the sentence level, in terms of a sentence accuracy, *SA*, or a sentence error rate, *SER*, see Table 6. In general, *SA* is lower than *WA*, because a single misrecognized word in a sentence impacts the *SA* parameter. It may, however, become higher than the word accuracy, especially when many single-word sentences are correctly recognized. The fact that *SER* and *SA* penalize a whole utterance when a single misrecognized word occurs has been pointed out by Strik, et al. [Strik-1] [Strik-2]; the problem can be circumvented with the parameters *NES* and *WES*, see Table 6. When utterances are not separated into sentences, all sentence-related metrics can also be calculated on an utterance instead of a sentence level.

Isolated word recognizers provide an output hypothesis for each input word or utterance. Input and output words can be directly compared and similar performance measures, as in the continuous recognition case, can be defined, omitting the insertions. Instead of the insertions, the number of false alarms in a time period can be counted, see van Leeuwen and Steeneken [van Leeuwen]. *WA* and *WER* can also be determined for keywords only, when the recognizer operates in a keyword-spotting mode.

Handwriting recognition is mostly measured by the recognition rate of symbols. Here, again the number of correctly determined symbols c_s, of substitutions s_s, of insertions i_s, and of deletions d_s per turn is counted.

For speech understanding assessment, two common approaches have to be distinguished. The first one is based on the classification of system answers to user questions into categories of correctly answered, partially correctly answered, incorrectly answered, or failed answers. The individual answer categories can be combined into measures which have been used in the US DARPA program, see Table 6. The second way is to classify the system's parsing capabilities, either in terms of correctly parsed utterances or of correctly identified AVPs. On the basis of the identified AVPs, global measures such as the concept accuracy, *CA*, the concept error rate, *CER*, or the understanding accuracy, *UA*, can be calculated. All parameters are listed in Table 6.

Table 6 – Speech-input-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. Meth.
WER, WA	word error rate, word accuracy	Percentage of words which have been correctly recognized, based on the orthographic form of the hypothesized and the (transcribed) reference utterance, and an alignment carried out with the help of the "sclite" algorithm, see [NIST SRST]. Designating n_w the overall number of words from all user utterances of a dialogue, and s_w, d_w and i_w the number of substituted, deleted and inserted words, respectively, then the word error rate and word accuracy can be determined as follows: $$WER = \frac{s_w + i_w + d_w}{n_w}$$ $$WA = 1 - \frac{s_w + i_w + d_w}{n_w} = 1 - WER$$ See [Simpson]; details on how these parameters can be calculated in case of isolated word recognition are given in [van Leeuwen].	Word	Instr./ expert.
SER, SA	sentence error rate, sentence accuracy	Percentage of entire sentences which have been correctly identified. Denoting n_s the total number of sentences, and s_s, i_s and d_s the number of substituted, inserted, and deleted sentences, respectively, then: $$SER = \frac{s_s + i_s + d_s}{n_s}$$ $$SA = 1 - \frac{s_s + i_s + d_s}{n_s} = 1 - SER$$ See [Simpson].	Turn	Instr./ expert.
NES	number of errors per sentence	Average number of recognition errors in a sentence. Being $s_w(k)$, $i_w(k)$ and $d_w(k)$ the number of substituted, inserted, and deleted words in sentence k, then $$NES(k) = s_w(k) + i_w(k) + d_w(k)$$ The average *NES* can be calculated as follows: $$NES = \frac{\sum_{k=1}^{\#\,user\ turns} NES(k)}{\#\,user\ turns} = \frac{WER \cdot \#\,user\ words}{\#\,user\ turns}$$ See [Strik-1].	Turn	Instr./ expert.

Table 6 – Speech-input-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. Meth.
WES	word error per sentence	Related to *NES*, but normalized to the number of words in sentence *k*, *w*(*k*): $$WES(k) = \frac{NES(k)}{w(k)}$$ The average *WES* can be calculated as follows: $$WES = \frac{\sum_{k=1}^{\#user\,turns} WES(k)}{\#user\,turns}$$ See [Strik-1].	Word	Instr./ expert.
AN:CO, AN:IC, AN:PA, AN:FA, %AN:CO, %AN:IN, %AN:PA, %AN:FA	number or percentage of correct/ incorrect/ partially correct/ failed system answers	Overall number or percentage of questions from the user which are • correctly (*AN:CO*) • incorrectly (*AN:IC*) • partially correctly (*AN:PA*) • not at all (*AN:FA*) answered by the system, per dialogue, see [Polifroni] [Goodine] [Hirschman].	Turn	Expert.
$DARPA_s$, $DARPA_{me}$	DARPA score, DARPA modified error	Measures according to the DARPA speech understanding initiative, modified by Skowronek [Skowronek] [Möller] to account for partially correct answers: $$DARPA_s = \frac{AN:CO - AN:IC}{\#user\,questions}$$ $$DARPA_{me} = \frac{AN:FA + 2\cdot(AN:IC + AN:PA)}{\#user\,questions}$$ See [Polifroni] [Goodine] [Skowronek].	Turn	Expert.
PA:CO, PA:PA, PA:IC, %PA:CO, %PA:PA, %PA:IC	number of correctly/ partially correctly/ incorrectly parsed user utterances	Evaluation of the number of concepts (attribute-value pairs, AVPs) in an utterance which have been extracted by the system: • *PA:CO*: All concepts of a user utterance have been correctly understood by the system. • *PA:PA*: Not all but at least one concept of a user utterance has been correctly understood by the system. • *PA:IC*: No concept of a user utterance has been correctly understood by the system. Expressed as the overall number or percentage of user utterances in a dialogue which have been parsed correctly/partially correctly/incorrectly [Danieli].	Turn	Expert.

Table 6 – Speech-input-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. Meth.
CA, CER	concept accuracy, concept error rate	Percentage of correctly understood semantic units, per dialogue. Concepts are defined as attribute-value pairs (AVPs), with n_{AVP} the total number of AVPs, and s_{AVP}, i_{AVP}, and d_{AVP} the number of substituted, inserted, and deleted AVPs. The concept accuracy and the concept error rate can then be determined as follows: $$CA = 1 - \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$$ $$CER = \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$$ See [Gerbino] [Simpson] [Boros] [Billi].	Turn	Expert.
UA	understanding accuracy	Percentage of user utterances in which all semantic units (AVPs) have been correctly extracted: $$UA = \frac{PA:CO}{\#\, user\ turns}$$ See [Zue].	Turn	Expert.

Gesture

Most input modalities, apart from speech and handwriting, can be roughly defined as gestural input. In GUI-based interactions, for example, the user selects buttons or items from a drop-down list, etc. by pointing gestures. And again, the number of correctly determined gestures c_G, of substitutions s_G, of insertions i_G, and of deletions d_G per turn is counted. On this basis, metrics comparable to the ones given in Table 6 can be calculated.

8.6 Output-related parameters

Output-modality appropriateness

As explained above for input modalities, multimodal systems may offer a set of output modalities that may be used sequentially, simultaneously, or compositely. Depending on the content, the environment, and the user, it can be determined (for example, guided by modality properties as described in [Bernsen-2]) if the output modalities are appropriate for every given turn. This can be annotated per modality or, in the case of composite output, for the multimodal output as a whole. In the first case, each modality can be appropriate or inappropriate. In the second case, the multimodal output can be appropriate, partially appropriate, or inappropriate.

Output-modality synchrony

For multimodal systems, the synchrony of simultaneous or composite output can be measured by the lag of time between corresponding modalities in milliseconds or by the overall number of times different output modalities are asynchronous.

Table 7 – Output-related interaction parameters

Abbr.	Name	Definition	Int. level	Meas. meth.
OMA:AP, *OMA:PA*, *OMA:IA*, *%OMA:AP*, *%OMA:PA*, *%OMA:IA*	output modality appropriateness	Overall number or percentage of chosen output modalities which are judged to be appropriate in their immediate dialogue context. Determined by labelling system output according to whether they violate one or more of Bernsen's modality properties: • *OMA:AP*: Appropriate. • *OMA:PA*: Partially appropriate. • *OMA:IA*: Inappropriate.	Turn	Expert.
LT	lag of time	Overall lag of time between corresponding modalities, in ms.		
# AE	number of asynchronous events	Overall number of times different output modalities are asynchronous.		

8.7 Further parameters

The majority of interaction parameters listed in the tables describes the behaviour of the system, which is obvious because it is the system and service quality which is of interest. In addition to these, user-related parameters can be defined. They are specific to the test user group, but may nevertheless be closely related to quality features perceived by the user.

9 Interpretation of interaction parameter values

Although interaction parameters, such as those defined in this supplement, are important for system design, optimization, and maintenance, they are not directly linked to the quality which is perceived by the human user. Consequently, the collection of interaction parameters should be complemented by a collection of user judgements on different quality aspects. Only in this way can valid information on the quality of services, which are based on multimodal dialogue systems, be obtained. Subjective evaluation methods for MDS are still under discussion in ITU-T.

An interpretation of interaction parameter values may be based on experimental findings which are, however, often specific to the considered system or service. As an example, an increased number of time-out prompts may indicate that the user does not know what to say at specific points in a dialogue, or that he/she is confused about system actions [Walker-1]. Increasing barge-in attempts may simply reflect that the user learned that it is possible to interrupt the system. In contrast, a reduced number may equally indicate that the user does not know what to say to the system. Lengthy user utterances may result from a large amount of initiative attributed to the user. A decrease of meta-communication-related parameter values (especially of user-initiated meta-communication) can be expected to increase system robustness, dialogue smoothness, and communication efficiency [Bernsen-1]. A high number of modality changes may indicate that the user profits from the advantages of multimodality. A high number of inappropriate input modality usage might point to a certain preference of the user or imply that the option to use, or the use itself, of the more appropriate modalities, is not clear to the user.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
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	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Terminals and subjective and objective assessment methods**
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, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2011