ETSI ES 202 076 V2.1.1 (2009-08)

ETSI Standard

Human Factors (HF); User Interfaces; Generic spoken command vocabulary for ICT devices and services

Reference

RES/HF-00081

Keywords

ICT, interface, speech, telephony, voice, user

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

Individual copies of the present document can be downloaded from:
http://www.etsi.org

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
http://portal.etsi.org/tb/status/status.asp

If you find errors in the present document, please send your comment to one of the following services:
http://portal.etsi.org/chaircor/ETSI_support.asp

Copyright Notification

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://webapp.etsi.org/IPR/home.asp).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This ETSI Standard (ES) has been produced by ETSI Technical Committee Human Factors (HF).

The work has been conducted in collaboration with industry. The present document is based upon user testing, empirical data, phonetic discriminability analysis, expert knowledge, and an industry-consultation and consensus process, aimed at a quick uptake and the widest possible support in product implementations to come.

Intended readers of the present document are:

- terminal manufacturers;
- service providers;
- network operators;
- manufacturers of multilingual speech recognizers;
- standards developers;
- software and user interface developers.

Introduction

Telecommunications, converging with information processing, and intersecting with mobility and the internet, are leading to the development of new interactive applications and services, offering global access.

A technology enabling a natural user interaction with these (often complex) systems and services is speech recognition. In recent years, speech recognition has become commercially viable in off-the-shelf ICT (Information and Communication Technology) devices and services. Just as the graphical user interface changed the way we interact with personal computers, so voice user interfaces are changing the way we interact with ICT devices and services.

Voice is fundamental to human communication and forms an important channel for universal access to ICT services. Voice user interfaces are a terminal, display and potentially location-independent user interface technology, enabled by speech recognition technologies. In order to simplify the user's learning and facilitate reuse of knowledge for the control of different applications and devices, it is desirable to standardize voice commands for the most common and generic functions. This standardization activity also meets one of the most important principles of the *e*Europe 2005 Action Plan; that of design for all. This theme has been continued by the new EU initiative; the i2010 Action Plan. This will help ensure that those with special needs such as elderly people, people with visual and other impairments, as well as young children will benefit from a generic spoken command vocabulary. As the standard necessarily addresses speech input it is recommended that the users of the present document provide some form of guidance for those end users who may have a speech impediment.

The present document is a timely contribution to enable the deployment of speech recognition in services and devices, offering multi-lingual voice user interfaces. Thereby it will minimize learning effort, facilitate knowledge transfer and develop user trust. Uniformity in the basic spoken commands improves the overall usability of the entire interactive environment, which becomes increasingly important in a world of ubiquitous devices and services using speech recognition.

The minimum generic set of spoken commands in the present document has been developed with a combined methodology, including the collection of data from native speakers of the 30 languages covered by the present document (see annex A for details). Therefore, it supports developers of ICT devices and services, leading to quicker, more consistent, cheaper, and better user interface development.

The work is aligned with, and co-funded by, the European Commission's initiative ***eEurope***, a programme for inclusive deployment of new, important, consumer-oriented technologies, opening up global access to communications and other new technologies, for all [2].

1 Scope

The present document specifies a minimum set of spoken commands required to control the generic and common functions of ICT devices and services that use speaker-independent speech recognition. It specifies the necessary and most common vocabularies for voice commands to be supported by ICT devices and services.

The present document is applicable to the functions required for user interface navigation, call handling, the control of and navigation in media, and management of device and service settings.

The present document specifies commands for the official languages (at the time of publication) of the European Union (EU) and the European Free Trade Association (EFTA) countries, and for Russian. The standard addresses Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Macedonian, Maltese, Norwegian, Polish, Portuguese, Raeto-Romance, Romanian, Russian, Slovak, Slovene, Spanish, Swedish, and Turkish [4]. Therefore, this updates the existing standard, ES 202 076 [1], which covers only the five languages with the largest number of native speakers in the European Union: English, French, German, Italian and Spanish. The present document does not cover dialects with the exception of Norwegian and Raeto Romance both of which have established dialects. All languages are addressed in "Received Pronunciation".

The present document does not cover dialogue design issues, the full range of supplementary telecommunications services, performance-related issues or speech output. Alphanumeric characters and symbols are not covered with the exception of single digits and language-specific reference to two recurring digits (e.g. "Double Two").

2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific.

- For a specific reference, subsequent revisions do not apply.
- Non-specific reference may be made only to a complete document or a part thereof and only in the following cases:
  - if it is accepted that it will be possible to use all future changes of the referenced document for the purposes of the referring document;
  - for informative references.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long term validity.

2.1 Normative references

The following referenced documents are indispensable for the application of the present document. For dated references, only the edition cited applies. For non-specific references, the latest edition of the referenced document (including any amendments) applies.

[1] ETSI ES 202 076 (V1.1.2): "Human Factors (HF) ; User Interfaces; Generic spoken command vocabulary for ICT devices and services".

[2] i2010 - A European Information Society for growth and employment.

NOTE: Available at http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm.

[3] ITU-T Recommendation I.210 (1993): "principles of telecommunications services supported by an ISDN and the means to describe them".

[4] Languages of Europe - The Official EU languages.

NOTE: Available at http://ec.europa.eu/education/policies/lang/languages/index_en.html.

[5] ISO 9241-11 (1998): "Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: guidance on usability".

2.2 Informative references

The following referenced documents are not essential to the use of the present document but they assist the user with regard to a particular subject area. For non-specific references, the latest version of the referenced document (including any amendments) applies.

[i.1] ETSI EG 201 013: "Human Factors (HF); Definitions, abbreviations and symbols".

[i.2] ETSI TR 102 068: "Human Factors (HF); Requirements for assistive technology devices in ICT".

[i.3] ETSI EG 202 048: "Human Factors (HF); Guidelines on the multimodality of icons, symbols and pictograms".

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in EG 201 013 [i.1] and the following apply:

basic command: employed frequently across a wide range of applications

design for all: design of products to be usable by all people, to the greatest extent possible, without the need for specialized adaptation

dialogue: series of exchanges between the user and a system

function: abstract concept of a particular use of or operation in a device or service

hot word: See keyword.

ICT devices and services: devices or services for processing information and/or supporting communication, which have an interface to communicate with a user

impairment: reduction or loss of psychological, physiological or anatomical function or structure of a user (environmental included)

keyword: word that the speech recognition system is looking for in word spotting mode

magic word: See keyword.

menu: list of choices from which a selection can be made

NOTE: A menu dialogue offers a user a series of lists of choices from which a series of selections can be made. The result from any one selection may be another menu.

phonetic discriminability: ability to discriminate between words based on the analysis of their constituent phones

spoken command: verbal or other auditory dialogue format which enables the user to input commands to control a device or service

supplementary service: additional service that modifies or supplements a basic telecommunication service

NOTE: Consequently, it cannot be offered to a customer as a stand-alone service; it has to be offered in association with a basic telecommunication service. The same supplementary service may be common to a number of basic telecommunication services. See ITU-T Recommendation I.210 [3].

usability: effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments (see ISO 9241-11 [5])

user: person who interacts with a product (see ISO 9241-11 [5])

user interface: elements of a product used to control it and receive information about its status, and the interaction that enables the user to use it for its intended purpose

user requirements: requirements made by users, based on their needs and capabilities, in order to make use of a product in the easiest, safest, most efficient and most secure way

word spotting mode: special state of the recognition system in which no speech is recognized or processed other than a limited set of keywords

NOTE: A typical usage is in a dormant state of the speech recognizer, where issuing a "wake up" command (also known as hot-word or keyword) can reactivate speech functionality.

3.2 Abbreviations

For the purposes of the present document, the following abbreviations apply:

ASR	Automatic Speech Recognition
CWD	Carefully Worded Description
EFTA	European Free Trade Association
EU	European Union
GPS	Global Positioning System
ICT	Information and Communication Technology
NLA	Native Language Assistant
UCU	University College Utrecht

4 User requirements

Intended *users* of the present document are those designing, developing, implementing and deploying ICT devices and services with a speech user interface.

Intended *end users* mentioned in the present document are people who use ICT devices and services with a speech interface, ranging from first time users to experienced power users.

Uniformity in the interactive elements increases the transfer of learning between different devices and services. Such knowledge transfer becomes even more important in a world of ubiquitous devices and services using speech recognition technology. In particular standardized commands improve the overall usability of the entire interactive environment. Use of the generic vocabulary of spoken commands in the present document for the development of ICT devices and services will enable end users to reapply knowledge and experience.

A generic spoken command vocabulary will particularly benefit some end users with temporary or permanent additional needs, such as those with literacy difficulties, people with visual or cognitive impairments, those with an impaired ability to perceive tactile stimuli, and people with limited dexterity.

For further guidance, including specifics of user impairments and resulting disabilities, assistive technologies, design for all and multi-modal interfaces, see TR 102 068 [i.2] and EG 202 048 [i.3].

Ideally, a spoken command vocabulary should be intuitive, easy to learn, memorable, natural, and unambiguous. A well-designed speech interface should:

- have a shallow learning curve;
- execute most common tasks;
- the ability to handle the vagaries of speech recognizers in a reliable and predictable way, maximizing the user experience.

Adequate feedback should be provided to users indicating, where applicable, that a command cannot be executed when requested. Three examples are:

- When a function is not supported.
- When the function is currently not available.
- When the command is not understood.

5 Method

5.1 General

In order to meet the requirements stated in clause 4, where the standard is designed for a wide range of end users, an empirical method has been employed for the elicitation and validation of potential voice commands. Native speakers of the 30 languages were sampled for this data collection. The previous standard used an online method of data collection where respondents were asked to complete a questionnaire. This worked well for the five most frequently spoken languages of the EU. However, the extension of the standard covers countries where internet penetration is relatively low and online questionnaires for these countries would not yield a representative sample of users for the purposes of inclusion.

In addition to elicitation and validation, a procedure of phonetic discriminability has been applied to the candidate commands to ensure minimal confusion with commands that are likely to be simultaneously available.

The employed method consists of three phases:

- Phase 1: Elicitation of command candidates;
- Phase 2: Validation of command candidates;
- Phase 3: Phonetic discriminability.

These phases are outlined here. More detailed descriptions of each phase can be found in annex A.

5.2 Elicitation of command candidates

In this phase, a sample of native speakers representing three age groups, aiming for an equal distribution of men and women, were invited to take part in an interview on voice commands. At this stage, they were given some general background to the aims of the study in order to inform them of the aims of the study prior to gaining their consent to participating in the research. In most cases the interview was conducted by telephone but, in a small number of cases, an interview was conducted with interviewer and interviewee sitting back to back in order to prevent artefacts based on the interviewer's reactions. The interviewer, or Native Language Assistant (NLA), was also always a native or near-native speaker who also carried out translations and transcriptions from documents in the original English and conducted analyses. They read out, for each command, a phrase describing the function of the device or service, known as the Carefully Worded Description (CWD), without mentioning any of the most likely resulting terms. The interviewees were then asked to name the term or terms they would find most suitable as a command in the context of a spoken-command supported device or service.

EXAMPLE: The carefully worded description used for describing the supplementary service "Call deflection" was: "You hear the phone ring at a time when you do not want to speak to anyone. You want the connection to be passed on to another name or number instead. What command would you give before saying this name or number?".

From this process a number of different alternative command candidates were collected. The lists of terms were then processed in order to reduce the number of morphological forms, e.g. infinitive or imperative, singular or plural, formal or informal addressing. The data were also checked for typological errors and answers which did not reflect the function implied by the carefully worded descriptions. The resulting terms were ordered according to the percentage of participants who had named them, and the most frequently chosen terms were used as input to the validation phase.

5.3 Validation of command candidates

In identifying the appropriate spoken commands it is not sufficient to conduct elicitation alone. It was also necessary to rank the proposed terms in order to provide a degree of validation. Therefore, validation interviews were set up and carried out in a similar way to elicitation interviews where the candidate commands were ranked in order of preference by the participants (see clause A.2). The top-ranked commands were then put forward to the phonetic discriminability phase.

The method described here was applied to the majority of the languages. However, it became clear that this method was an unnecessary use of resources as the same result could be obtained by subjecting the results from discrimination to expert analysis. Therefore, (see clause A.2), expert analysis was applied to those languages which had not undergone validation, namely: Estonian, Greek, Icelandic, Latvian, Maltese, Norwegian, Portuguese, Raeto-Romance, Swedish, and Turkish to identify the spoken commands which were chosen for phase 3, phonetic discriminability. The experts comprised a combination of: the NLAs, industry experts, linguistic and cultural representatives from the countries involved, and Human Factors experts.

5.4 Phonetic discriminability

Whilst the previous two steps have provided a user-centric approach to the selection of command words, it is still important to address technology issues.

EXAMPLE: A selection of words may be chosen as a result of the previous two phases that have a high level of agreement across the user group.

However, if this selection gives rise to a high degree of confusability in the speech recognizer, between words which are available for use in the same context, then the overall goal of usability is nullified. Therefore, discriminability analysis was carried out to ensure that command words that are likely to be active simultaneously in a dialogue context can be recognized correctly by the speech recognition system.

The approach consisted of the following steps:

a) Commands were clustered according to those which would be simultaneously available, e.g. all commands for functions related to the handling of phone calls.

b) For each context, the top three commands from validation were assessed by native-language experts with respect to their sounds and not to their orthographic forms. Commands were listed as potentially phonetically confusable if:

- they share the same initial consonant or consonant cluster;
- they share similar stressed vowels;
- they rhyme;
- they are of equal length.

c) Commands that give rise to possible phonetic confusion were collated.

d) An alternative for one of the command words was chosen, with minimum repercussion with respect to the ranking of candidates.

5.5 Final command definition

The final pass on the resulting command set was performed by submitting the results to a number of different groups for verification. These were:

- Educated native speakers to ensure consistency within the entire language set in terms of morphological and other characteristics.
- The NLAs, who were all native speakers of the languages they assisted with.
- Cultural and linguistic institutes of each of the languages represented in the standard.

- The industry reference group. This is a body of experts from industry, such as service providers and handset manufacturers, who would be responsible for the implementation of the standard in some or all of the countries involved.
- Experts in the design of ICT products and services for all.

6 List of commands

6.1 Principles of use

The spoken commands specified in the present document are divided into the following categories:

1) basic commands;

2) digits;

3) communication commands;

4) commands for the control of and navigation in media;

5) commands for device settings.

For the present document, the following principles of use in implementations apply, assuming a speech recognition user interface is provided:

1) The ICT device or service shall support all the commands specified in the present document if the corresponding functionality is implemented.

2) If a function as defined in the present document is not supported by the ICT device or service, the corresponding command should still be accepted as user input and guidance information should be provided to the user.

3) The commands specified in the present document can be concatenated into more complex expressions (e.g. "Call - Paul - Home", or "Divert to - Five - Seven - Nine").

4) In addition to the commands specified in the present document, alternative and additional commands may be offered by the device and service provider. However, additional commands should be tested for phonetic discriminability with other commands available in the same context.

5) One word which was suggested for inclusion in the standard is "Select". This allows users to choose an item from a menu. However, the suggestion came too late for the data collection exercise. This word may be the subject of an extension to the present document but, in the mean time, command 1.1 in table 1.a ("confirm operation") may provide a suitable command.

6) For some languages, one command is used for more than one function (e.g. 3.1 and 3.2). However, in these cases, the command should be disambiguated by their different contexts.

7) In some languages functions are covered by one command, in other languages alternative commands exist for those same functions. This is a direct result of the empirical data collection and subsequent analysis.

For clarity where there is more than one command for a function, these commands have been separated by commas and the first word of a command starts with an upper case letter.

8) For commands for emergency services (3.7) only the relevant words in each language are given. The spoken commands for the digits 112 are already specified in clause 6.3. In addition, if a user wanted to say "Call 112" or "Dial 112", the relevant word for "Call" or "Dial" is also specified in clause 6.4.

9) Two of the official languages of EFTA member countries Norway and Switzerland are represented by more than one variant, namely Bokmål and Nynorsk for Norwegian (Riksmål and Høgnorsk have not been considered), and Ladin, Surmiran, Sursilvan, and Rumantsch Grischun for Raeto-Romance (Sutsilvan, Putér, and Vallader have not been considered). Which of these variants is represented in a given command is indicated by indices and footnotes in the respective tables.

6.2 Basic commands

Basic commands are employed frequently across a wide range of applications but they maintain the same effect, irrespective of the (dialogue) context in which they are executed. The meaning of each basic command is explained in table 1a, and the language-specific versions of the basic commands in the 30 languages are presented in tables 1b through 1g.

EXAMPLE: A user is unfamiliar with a new spoken-command system. She activates it ("Wake-up") and requests it to list the available commands ("Options"). After exploring some of the supported functionality, she decides to return to the main menu of the command tree ("Main menu") to navigate to a specific application. Once taking the wrong menu tree branch, the user returns to the previous menu-tree position ("Go back"). The application she then activates requires some initial input from her by asking some questions that are answered either affirmatively ("Yes") or negatively ("No"). In one case, she did not understand the question properly and asks the system to repeat it ("Repeat"). Following this, she leaves the voice-command system for a break ("Standby"). After returning to the spoken-command system, the user activates the help system ("Help") and receives some voice-based explanation, but decides that this is not helpful ("Stop") and asks to be connected to a human operator ("Operator"). After having completed her activities, she shuts down the system ("Goodbye").

Table 1a: Basic commands

Index	ICT device/service function	Explanation
1.1	Confirm operation	Positive confirmation
1.2	Reject operation	Negative confirmation
1.3	Wake-up the speech recognizer (ICT device or service in word spotting mode)	ASR ignores all speech input, except a wake-up command (hot-word, magic word or keyword). When this command is detected, the recognizer switches to a larger active vocabulary, determined by the dialogue design
1.4	Enter idle mode	Put the service into monitoring mode for a wake-up command
1.5	Terminate service	Get off line, end session
1.6	Help	Provide context-dependent explanations and guidance (may provide more detailed help on repetition of the command)
1.7	Transfer to human operator	Leave the speech recognition mode and transfer to a human attendant, an operator, in telecommunications-specific contexts. This command should also be used when offering relay services
1.8	Go to top level of service	Leave current function, go to main menu or application
1.9	List commands and/or functions	Request for listing of available commands (optionally with their functionality)
1.10	Cancel current operation	Immediately abort ongoing operation (e.g. during the (long) playback of a recorded message)
1.11	Go back to previous node or menu	Navigate backwards in a dialogue structure (can also be used to cancel a forced choice operation)
1.12	Read prompt again	Repetition of the last acoustic feedback message

Table 1b: Basic commands (Bulgarian, Croatian, Czech, Danish, Dutch)

Index	ICT device/service function	Bulgarian	Croatian	Czech	Danish	Dutch
1.1	Confirm operation	Да	Da, Izvrši	Ano	Ja, Udfør	Ja
1.2	Reject operation	Не	Ne, Odustani	Ne	Nej, Annuller	Nee
1.3	Wake-up the speech recognizer (ICT device or service in word spotting mode)	Активирай	Aktiviraj	Vstávat	Aktiver	Activeren