

## STANDARDISING SPOKEN COMMANDS FOR MOBILE DEVICES AND SERVICES:

*The Revision of ES 202 076 Based on Empirical Data Collection*

MARTIN BÖCKER<sup>1</sup>, HELGE HÜTTENRAUCH<sup>2</sup>, ROSEMARY  
ORR<sup>3</sup>, FRANÇOISE PETERSEN<sup>4</sup> AND MIKE TATE<sup>5</sup> - EUROPEAN  
TELECOMMUNICATIONS STANDARDS INSTITUTE (ETSI) STF  
326<sup>6</sup>

<sup>1</sup>*Böcker & Schneider GbR, Germany {boecker@humanfactors.de},*  
<sup>2</sup>*vonniman consulting, Sweden,* <sup>3</sup>*UCU, The Netherlands,* <sup>4</sup>*Apica,*  
*France,* <sup>5</sup>*Ontologic, Research Ltd., UK (STF 326 Leader),* <sup>6</sup>*Spoken-*  
*Commands@etsi.org*

**Abstract.** The current state of the art in voice-recognition technology is reaching a level of maturity that suggests that sophisticated voice-based user-interface technologies will soon be available also in mobile devices the computational capacity of which matches that of the personal computer of only a few years ago. As voice commands for mobile devices and services have so far been primarily limited to voice dialling and the voice activation of individual menu items, there is currently a window of opportunity to standardise voice commands for basic functions of mobile devices and services. ETSI, the European Telecommunications Standards Institute, has formed a Specialist Task Force (STF) to extend the existing ETSI standard ES 202 076 (published in 2002) on voice commands for mobile services and devices. The previously published standard covers five major European languages, whereas its revised version will extend the number of languages covered to 30. This paper gives an overview of the commands and languages covered, and it outlines the approach for the development of the voice commands (elicitation, validation and phonetic discrimination). In addition, first results, as available at the time of writing, are presented.

### 1. Introduction

Many modern ICT applications support speech as a user-interface modality. Although the speech interface is not the most common form of user interface,

it is the most natural form of human communication, and there are a number of user groups for whom the speech-enabled user interface is particularly valuable or even crucial. Among these are the “hands-busy” users, such as people working with their hands, or those driving cars, as well as users with special needs, such visually-impaired users, users with reduced ability to perceive tactile stimuli, and users with limited dexterity. Furthermore, voice control is one of the key user-interface technologies for ambient-intelligence applications such as smart homes.

## **2. Standardising voice commands**

As technology continues to spread to all sections of society, the proportion of users who can benefit from speech-driven applications is increasing. Before they can effectively make use of a speech interface, users have to learn a specialized vocabulary. There is, however, a danger that different manufacturers and service providers will develop their own sets of commands, which would lead to the undesirable consequence of users having to learn different commands when moving between devices and services of different manufacturers or when they upgrade from one generation of devices or services to the next. In addition, there is a risk that a person who is used to saying a particular command word for a specific function within a device may invoke a different functionality when using a device from another manufacturer.

The telecommunications industry is beginning to acknowledge that uniformity in basic user-interface elements across technologies and brands (see e.g. [1 and 2]) has the potential of increasing the transfer of user learning, and therefore also of growing revenue through increased usage. The standardisation of a vocabulary for spoken interaction with technology has a similar potential of facilitating the re-application of knowledge and, thereby, the uptake and usage of that new technology. For this reason, ETSI published European Standard ES 202 076 “Generic spoken command vocabulary for ICT devices and services” [3] in 2002 which specifies a set of spoken commands for five of the largest European languages, in terms of numbers of speakers in 2002, i.e. English, French, German, Italian, and Spanish. As ES 202 076 was beginning to be implemented by various manufacturers, ETSI set up a specialists task force (STF) to extend the standard to cover more languages. That work is being co-funded by EU/EFTA within the i2010 Policy Framework [4].

### 3. Scope of revised ES 202 076

#### 3.1. LANGUAGES COVERED

The revised version of the standard will cover the official languages of the European Union and of EFTA (European Free Trade Association) as well as those of countries having the official status of applicants to the European Union. In addition, Russian has been included as the most important European language in terms of the number of native speakers.

TABLE 1: Languages covered by the standard.

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	Slovakian	Slovenian
Spanish	Swedish	Turkish

#### 3.2. COMMANDS COVERED BY THE STANDARD

The commands covered by the standard were selected as representing those needed for some of the most likely use cases encountered by mobile users (see Table 2). They are organised in clusters of commands for different sets of functions: “Basic commands” are expected to be used in most applications and include meta-commands for controlling the voice application itself; “Digits” are important *inter alia* for communications applications such as dialling; “Communications commands” deal with the handling of calls, including some for setting up and diverting calls as well as in-call functions such as switching between calls and setting up a conference call; the group of “Commands for the control of and navigation in media” allow the user to control media presentation and to select and modify (media) items; and the “Commands for device and service settings” deal with the setting up of device and service parameters. The selection of commands covered has not changed since the 2002 version of the standard. While it would have been desirable to extend the scope of the standard in terms of the command clusters covered, extensions of the command set such as, e.g. GPS-based navigation, are recommended to be addressed in later revisions of the standard.

TABLE 2: Commands covered by the standard.

Basic commands	
Confirm operation	Reject operation
Wake-up recognizer	Enter idle mode
Terminate service	Help
Transfer to human operator	Go to top level of service
List commands and/or functions	Cancel current operation
Go back to previous node or menu	Read prompt again
Digits	
Digits 0 to 9	Next digit repeated twice (“Double O”)
Enter international access code	
Communications Commands	
Initiate digit dialling sequence	Dial a number or name
Home phone number (location)	Work phone number (location)
Mobile phone number (location)	Car phone number (location)
Personal number (attribute)	Make a call to the emergency services
Redial last dialled number	Set up a call-back to a called number
Accept incoming call	Reject incoming call
Forward incoming call	Set up a call diversion
Transfer an ongoing call	Put call on hold
Switch between two calls (hook flash)	Set up a conference call
Commands for the control of and navigation in media	
Play a recording	Start a recording
Stop temporarily	Resume interrupted playback
Stop playing a recording	Move forward faster than play
Move backward	Go to previous item
Go to next item	Provide more information about item
Modify item	Store item
Remove item	Respond to item
Forward item	Create new item
Send item	Move item to a new location
Reapply the undone action	Reverse the previous action
Commands for device and service settings	
List networks	Increase the volume
Decrease the volume	Silent mode
Re-activate the audio output	Silence the loudspeaker
Reactivate the loudspeaker	Silence the microphone
Re-activate the microphone	Activate vibrating alert
Deactivate vibrating alert	Summary of current device status
Change profile (pre-stored settings)	

## 4. Method

### 4.1 GENERAL APPROACH

The method of establishing generic vocabularies for this extended standard differs in some important respects from the general principles described in the 2002 version of the standard [3]. In particular, the approach chosen for the revised standard consists of the following three phases:

- Elicitation: candidate commands generated spontaneously by participants (native speakers);
- Validation: ranking of the most successful commands resulting from the Elicitation phase by participants (native speakers); and
- Phonetic discriminability testing; a process to ensure that the selected commands do not have phonetic properties that will confuse the voice-recognising device.

### 4.2. PROCEDURE

#### 4.2.1 *Elicitation Procedure*

The main purpose of this phase was to collect voice commands spontaneously generated by participants. In order to prevent methodological artefacts, the briefing for the participants was defined very carefully in order not to influence or prime them for certain responses. In particular, the description of the command in question was not allowed to contain the commands most likely to be named by the participants. This was achieved by phrasing carefully-worded descriptions (CWDs) for each of the 63 functions covered. These descriptions were chosen such that they avoided potential command words. After having listened to a CWD, the participants were asked whether they had understood the description. They were then asked to name up to three proposals of what would be a suitable command name for it (for this purpose, a fictitious voice-command system ('Speak to me') was introduced). Following the completion of the elicitation interview series, each team of interviewers (one for each language) standardised the raw data in terms of bringing them into an appropriate morphological form for that language, e.g. infinitive or imperative).

#### 4.2.2 *Validation Procedure*

The most successful commands emerging from the descriptive statistical analysis (frequencies) of the Elicitation phase data were entered as input into the Validation Procedure (for each function, all commands that were included until 85% of the responses were reached). The procedure for the Validation phase resembled that of the Elicitation phase in that participants were briefed about the general purpose of the study and about the 'Speak to me'-scenario.

The top-candidate term from the elicitation phase was presented to the participant, who had to describe the functionality they associate with that command. If the interviewee did not succeed, then the second candidate was then presented. This was followed for each of the commands by the interviewer reading out the successful command candidates resulting from the Elicitation phase plus one dummy item, all presented in random order. The purpose of the dummy item was to ensure that the candidates understood the CWD. The participants' task was to rank the commands in terms of their suitability as a voice command for the function in question.

#### *4.2.3 The Elicitation and Validation interviews*

The interviews took place either with one interviewer and one participant in the same room sitting back to back (in order to avoid experimenter artefacts), or via telephone (in many instances using IP-telephony). The interviews began with up to three minutes of conversation between interviewer and participant, during which the purpose of the study and the procedure were explained. The interviewers also explained that they were going to change to a more formal and scripted behaviour and that they could only give limited help. Particular emphasis was laid on explaining that the interviewers were not "tested" in any way and that there were no "right" or "wrong" answers. During the interview, some socio-demographic data were collected related to the background and personal details of the participant, addressing in particular, age and experience with ICT applications. This was followed by the actual Elicitation or Validation interview items (see above).

#### *4.2.4 Phonetic discriminability testing*

The purpose of this test is to check the proposed commands for phonetic similarity with other commands that are likely to be active in the same situation, causing confusion in the Automatic Speech Recogniser (ASR). The number of incorrectly recognized commands can be reduced if the available words in a given context can be differentiated according to their acoustic properties. As it was not possible to conduct a recogniser field test or a pronunciation dictionary test (not for all of the 30 languages covered speech-recognition technology is available in sufficient maturity), a more pragmatic but effective approach was chosen which consisted of the following steps:

- a) Commands were clustered according to those which would most likely be simultaneously available (depending on implementations);
- b) Commands were listed as potentially phonetically confusable if they share the same initial consonant or consonant cluster, if they share similar stressed vowels, or if they rhyme;
- c) The collation of commands that give rise to possible phonetic confusion and the choice of alternative commands, with minimum repetition with respect to confidence ratings.

The final choice of commands will be based on results of the phonetic discriminability test and final expert judgement. The draft command set will also be circulated among national bodies for review. In addition, a group comprising representatives from relevant industries has been organised and will convene to discuss the standard prior to publication.

### 4.3. PARTICIPANTS

#### 4.3.1 *Interviewers*

The team interviewers taking part in the two empirical phases of the study consisted of members of staff and students of different university colleges as well as of members of the ETSI STF. A large number of the interviews were conducted by students of the University College of Liberal Arts at the University of Utrecht in the Netherlands. All interviewers were native speakers of the language spoken in the interviews. They were all briefed about the aims of the study and received a detailed training about the test procedure and the use of the software for collecting the participants' responses.

#### 4.3.2. *Participants*

In many instances, the participants were drawn from the social network of the interviewers. In many cases, the number of participants exceeded the target sample size of 30, for a very few cases, only a smaller number of participants could be recruited. In all cases, an equal distribution of the participant samples in terms of gender and age groups (15-25 years, 26-49 years, and 50 years or older) was intended. In total, more than 1.500 participants will have taken part in the Elicitation interviews and more than 800 in the Validation interviews after the final interviews will have been completed.

The participants were encouraged to take part in the study by informing them that they would be contributing to an international standard. In addition, participation relied, to some extent, on favours as many participants were acquaintances or family relations. Finally, the participants took part in a draw for a small prize. Wherever possible, the interviewer tried to ask the participants for help in finding further participants to take part in the study.

## 5. Results

At the time of writing, all of the interviews from the Elicitation phase (with very few exceptions) and a large number of the interviews from the Validation phase have been conducted. In addition, the method chosen for the phonetic discriminability testing has been piloted.

Already at this stage, a number of observations can be made about the expected results and the suitability of the chosen method. The participants tended to quickly understand the concept of circumscribing a function with

the carefully-worded description. Many participants did not have any personal experience with some of the functions covered. In some languages, the problem arose that one and the same command word was suggested for two functions (e.g. “Go to next item” and “Move forward faster than play”) that may well be active in the same use case. It was not always obvious what the appropriate morphological form for command words in a particular language are. Even though it is obviously desirable (at least from an implementation point of view) to standardise only one command word per function (and per language), many instances arose of two or more command words needing to be recommended.

## 6. Discussion

The implementation of the voice commands to be published in the revision of ETSI standard ES 202 076 will contribute to the uptake of services using speech as a user-interface modality, supporting all users by eliminating the need of learning separate command sets for devices and services from different manufacturers. The method chosen for the elicitation and validation of the commands to be recommended ensures that the command words to be learnt by users will in most cases seem to them both natural and intuitive.

## Acknowledgements

We acknowledge the co-operation of the experts from the ES 202 076 standard [3] in the preparation of this work, namely Bruno von Niman, Catriona Chaplin, David van Leeuwen, Lutz Groh, and Scott McGlashan. Also, the authors thank Ms. Dia Bene and Mr. Bernat Gonzalez as well as the members of staff of the participating universities and institutes for their enthusiastic support of the study. Finally, we wish to thank all participants from each language group who took part in the interviews.

## References

Comment:

ETSI publications can be downloaded from <http://pda.etsi.org/pda/queryform.asp>

- [1] ETSI EG 202 132 v1.1.1: 2004, Human Factors (HF); User Interfaces; Guidelines for generic user interface elements for mobile terminals and services.
- [2] ETSI ES 202 130 v2.1.2: 2007, Human Factors (HF); User Interfaces; Character repertoires, orderings and assignments to the 12-key telephone keypad (for European languages and other languages used in Europe).
- [3] ETSI ES 202 076 v1.1.2: 2002, Human Factors (HF); User Interfaces; Generic spoken command vocabulary for ICT devices and services.
- [4] i2010, A European information society for growth and employment, [online]: [http://ec.europa.eu/information\\_society/europe/i2010/index\\_en.htm](http://ec.europa.eu/information_society/europe/i2010/index_en.htm). (last accessed in January 2007).