

Standardization supporting cultural diversity: From 5 to 28 - Expanding the language coverage of the ETSI spoken command vocabulary standard

*Rosemary Orr¹, Lutz Groh¹, Helge Hüttenrauch¹,
Françoise Petersen¹, Michael Tate¹ and Bruno von Niman²*

¹ETSI Specialist Task Force QD

560, Rue des Lucioles, Sophia Antipolis, France

rorr@ucu.uu.nl

²Work Item Rapporteur

Abstract

The current ETSI Standard for spoken commands in ICT applications [ETSI ES 202 076] specifies user tested spoken commands for the five languages with the largest number of native speakers in the European Union, namely English, French, German, Italian and Spanish. Considering the imminent expansion of the EU and the strong presence of Russia in the European market, a revision of the ES has been initiated, in order to provide similar support for 28 languages, including those of the current EU states, the EFTA countries, the candidate EU states and Russia. The new ETSI Standard will enable the creation and use of interactive, accessible, multi-cultural and more modern online public and private services, making use of the widespread availability of fixed and mobile broadband access and infrastructure. Without the development of European or international standards and guidelines, many official European languages will not be supported in product implementations, and inadequate or inconsistent solutions based on proprietary standards will dominate product technology. In this paper, we will discuss the background to the standard along with alternative methods for choosing words in recognition devices, based on user needs and recogniser discriminability.

Key words: Automatic Speech Recognition (ASR), usability, transfer of knowledge, e-Europe, digital divide, accessibility, command and control, acoustic discrimination, phonetic discrimination, confidence rating

1. Introduction

Speech is a fundamental paradigm for human communication, and as such, has an important role in ensuring universal access to the services and benefits of communications technology. Automatic speech recognition (ASR) is a technology which enables this most natural user interaction with often complex ICT systems, devices and services.

In recent years, ASR has become commercially available and viable in off-the-shelf products and services. As consumer interest in ASR is increasing and more products include ASR in their implementations, it is of interest to producer and consumer alike, as well as society on the whole, to standardize the vocabulary used by the consumer.

The necessity for standardization has several motivations. Learning the vocabulary necessary to use a voice-driven technology requires some time on the part of the user. As long as the vocabulary can be used for multiple devices, this investment is well-spent, and ideally, the

knowledge built up in the familiarisation phase could be applied to all voice-driven devices. In order to maximise this transfer of knowledge, the most common and generic navigation, command and control, and editing vocabularies need to be standardized. As demonstrated already by the use of the current standard [ETSI ES 202 076] in various device and service implementations, uniformity in the basic interactive elements does increase the transfer of knowledge between devices and services using spoken commands and improves the overall usability of the interactive environment. Such transfer of knowledge becomes even more important in a world of ubiquitous devices and services using speech recognition.

The increased availability of ASR technology in various devices and services has raised the issue of support for multiple languages. Where English is often considered the *lingua franca* of Europe and many of the surrounding countries, the use of English in voice user interfaces limits the accessibility of this technology to very specific sections of the European market. The current ETSI Standard [ETSI ES 202 076], developed under the eEurope 2002 initiative, and funded by the EC/EFTA, defined vocabularies for the five most widely-spoken European languages, namely English, French, German, Italian and Spanish. This, however, can hardly be considered as covering the requirements of the local user on a Europe-wide basis.

The expansion of the EU has brought with it an increased sense of the importance of national or cultural identity for many nations, and the availability of *native language* user interfaces is experienced as helping to preserve and promote cultural diversity. The flourishing software localisation industry bears witness to the need to cater to users in their native languages, and there is every reason to suppose that, for voice-driven user interfaces, support in local languages will increase local usability. The availability of ASR technology in local languages allows for increased participation of populations with limited access to the languages defined in the current standard [ETSI ES 202 076].

The voice user interface has the added advantage of being a terminal-, display- and location-independent user interface technology. Users who have limited access to terminals and display, for example blind users, disabled users, or young children often also form populations with limited access to languages outside their native tongue, and again, access to technology for these groups may be substantially increased if local languages are supported.

With this in mind, we consider that the availability of a spoken command vocabulary in local languages would benefit not only the consumer but also society in general and its cultural diversity, taking another small step towards the achievement of the goals of “e-accessibility, combating the digital divide, stimulating quality of life, and encouraging participation” as set by the European Union in 2004 [5].

The aim therefore, of the proposed work is to enable users to reapply knowledge and previous experience between different devices and services and to control common functions by using a generic vocabulary of spoken commands. Furthermore, the accessibility of the voice user interface will be improved by extending standardization of vocabularies to more than just the five largest native languages in the European Union.

2. Objectives

Our work has the objective of developing an ETSI Standard to include all official EU and EFTA languages along with languages of the candidate EU states and Russian, which:

- enables Member States to take specific measures for disabled end-users in order to ensure access to and affordability of publicly available telephone services, including access to emergency services, directory enquiry services and directories through a

speech user interface, equivalent to that enjoyed by other end-user in accordance with the Universal Service Directive from 2002 [7]

- provides users with an additional modality for input of commands in situations where text input is inconvenient or even dangerous such as when driving a car
- supports the creation of an accessible, multi-cultural information society for all, as required by the *eEurope* Action Plan;
- facilitates the implementation of additional European languages, in addition to the five largest languages presently supported by many mass-market products and services deploying speech user interfaces

2.1 Taking previous standards and guidelines into consideration

The proposed work will be based on [ETSI ES 202 076, 2002], reapplying the methodology but covering 28 languages in total, including a revision of those previously developed. The work will also take into consideration the guidelines provided in [ETSI EG 202 132], developed under the *eEurope* 2002 Action Plan, specifying design guidelines for generic mobile user interface elements for mobile terminals and services. In addition, the guidelines and recommendations provided in [ETSI EG 202 191] will provide valuable guidance to the work and those who will later implement the command sets.

3. Methodology

The methodology consists of several distinct steps:

- *Spontaneous generation of potential command words*
The purpose of this step is to make an inventory of words that users would intuitively use, given the task that they want to complete.
- *Confidence rating of the found potential command words*
The purpose of this step is to ensure that the words found in the first step are considered likely to complete the task when a test subject is given the choice to use the word.
- *Phonetic discrimination*
The purpose of this step is to ensure that command words that can be active simultaneously in a dialogue context can be recognized correctly by the speech recognition system.

There are several ways of performing each step. In the following clauses we explain the methodologies in further detail.

3.1 Spontaneous generation of potential command words

In order to ensure that command words for speech recognition enabled devices and services are intuitive, some evidence must be gained as to which word(s) a user would use without prior training or experience. It is not trivial to find this out, because in order to get this kind of information the (potential) user is likely to be primed for certain words or phrases. For instance, if a test is set up where the task to be performed is explained to the test subjects, it is likely that the explanation contains some words that are candidate command words. If, on the other hand, dialogues of users of actual running systems are analysed, it is likely that the command words found in these dialogues are the words that the system designer has chosen and that the user has learnt to use.

Here we describe two methods that allow the collection of spontaneous command words from test subjects. In both methods, test subjects play a vital role. They must be recruited among people who understand the services for which the command words are sought, and are

familiar with the functionality, but are not actual users of speech-enabled implementations of such services. Firstly, for all services, the conceptual functionality to be supported must be determined and described. Secondly, for each of the functionalities, the intended functionality must be indicated to the test subject without their being primed for particular words.

3.1.1 Storyboard method

In this method, a professional artist makes a so-called storyboard, a set of illustrations or cartoons, for each function. The background is explained to the test subject and they are shown the illustration, then asked to give the command they would use in order to activate the shown functionality. This method has been described by MacDermid and Goldstein [MacDermid and Goldstein 1996]. The advantage of this method is that the same storyboard can be used for several different languages, as long as there are limited cultural constraints involved. The disadvantage is that some functionality might be very difficult to describe pictorially, and that there can be quite a lot of effort necessary from the artist.

3.1.2 Carefully worded descriptions method

In this method the functionalities are described textually in a paragraph of text, which is carefully constructed not to use any word, which might possibly be used as a command word. This method has been described by Guzman et al. [Guzman, S. J. et al, 2001]. The advantage of this method is that there is no need for a highly skilled professional for generating the textual descriptions. The disadvantages are that the descriptions may sometimes turn out to be very clumsily constructed in order to prevent using an obvious command word, and that this effort must be carried out in all languages in which one wants to conduct the inventory. Also, these must be carefully developed for each target language.

3.2 Confidence rating of command words

When the spontaneous words have been generated with a sufficient number of test subjects, a histogram of the word frequencies can be made. This histogram gives a lot of information about the variability in responses. Thus, it can be seen immediately what the most likely command words are. Integration of the sorted frequencies indicates what the coverage of spontaneous commands is, when the most frequent words are available for recognition.

For the confidence-rating test described, one can select the most frequent words that cover at least a given percentage (say, 80 %) of the spontaneously generated words. For instance, for "confirmation" this might be "yes," "sure" and "no problem," if these are the most frequently generated commands and cover 85 % of the responses for the functionality "confirmation".

The procedure described above does not guarantee that the command words imply the targeted functionality. For instance, for functionality "confirmation" the command phrase "why not" might have come up in the list of spontaneous commands, but reversibly it might not be obvious to a user that the command "why not" implies confirmation; it might suggest the inquiry of a reason. A confidence rating of command words tests how likely it is that the given command word implies the correct functionality.

Because the "spontaneous generation of words" test is an open response experiment, many different expressions for very similar command words can be obtained. Therefore, a manual check of the histogram may be necessary, where responses with the same the essential term are grouped together. Thus, for each function, a set of candidates for the confidence test can be obtained.

A way of measuring the confidence of command words is the following, after the work of Guzman [4]. A group of test subjects that are independent of the ones used to generate command words is presented the same data as in the first test, but are requested a different response. The test data can again be either from the "storyboard" or "descriptions" method. Instead of asking for a spontaneous command, the candidate commands from the first test are shown. For each of the commands, the test subjects are asked how confident they are that the command word will imply the functionality, on a 5-point scale. Instead of an explicit confidence measure, the subjects can also be asked to choose the command word(s) for which they have most confidence.

3.3 Phonetic discrimination

It is essential that command words are recognized correctly in a voice-enabled application. Although a system can ask for confirmation for certain not undoable operations (e.g. "delete subscription"), it is not acceptable if almost every command needs confirmation (e.g. "do you want to hear the next item? Please say yes or no"). For a given application or service there will be several contexts defined in which certain command words will be available, e.g. in the context of "confirmation question," words like "yes," and "no" will be active in the ASR vocabulary. The number of incorrectly recognized commands can be reduced if the available words in a given context are acoustically reasonably different. There are several ways to test the discriminability. We will assume that for the service or application the ASR contexts are well defined. For a given context, there will be a number of words active.

3.3.1 Recognizer field test

One can find out the acoustic discriminability by a field test with a real speech recognition system. This test gives realistic discriminability measures, but the results are sensitive to many chosen parameter settings such as type of recognition system (e.g. brand, speaker dependency, noise robustness) and test database (e.g. recorded speech samples versus live speech from test subjects, speaking style). For a project including 28 languages, in the absence of a multilingual recognizer, 28 different recognizers would be required.

The recording of test databases for voice commands requires quite a lot of effort, but there are several large databases for voice commands available. The test is conducted by preparing the ASR to recognize the required command set for each context, and then test each context with several instances of all the available commands within the context, uttered by many different test subjects. The *confusability* of a command word *A* with respect to an alternative command word *B* can be defined as the fraction of times and utterance of word *A* is recognized as word *B* by the recognition system. A confusion matrix for each context, containing the confusability of all active menu words with respect to each other, can indicate which command words pose particular problems to the recognition system. The *discriminability* of a set of command words is a measure that characterizes the whole confusion matrix. If the test database consists of recorded speech, the test can be repeated for another ASR system. This will give insight in the recognition system dependency of the discriminability results.

3.3.2 *Pronunciation dictionary test*

An alternative to a field test is the analysis of the acoustic realizations of the command words. This can be performed without collecting speech databases or test subjects, but the predictions are not validated. The only thing necessary is a *pronunciation dictionary*, a tool that is used often by speech recognition ICT device or service developers. A pronunciation dictionary consists of a lookup table of words in terms of their *phone sequences*, (phone is a separate unit of sound, similar to a phoneme in linguistics). For instance, the phone sequence for "yes" may be specified as the sequences "j eh s" or "j ea" (where we have introduced a Latin character readable phone symbols "j" "eh" "ea" and "s"). Typically, for a Western language, phone sets of 40 to 60 phones are defined for ASR systems.

The acoustic discriminability of two command words can be predicted on the basis of the phone sequences of the words. The number of different phones (order is important) might be called the first order prediction of the discriminability. For instance, in the context "start" "stop" the number of different phones is 2 for "stop" and 3 for "start". This is a relatively low number compared to the number of phones in the words, respectively 4 and 5. As a contrast, the context "begin" "end" has no phones/positions in common, so the number of different phones is 5 and 3, respectively.

A more elaborate scheme takes into account the confusion probability of two phones. For example, most ASR systems, as well as humans, have difficulty in distinguishing between "m" and "n". For any particular ASR system, these phone confusion probabilities may be measured, but this requires quite an elaborate test set-up of the ASR system. If this information is not available, the phones in a language might be grouped, and the discriminability can be measured in terms of the different phone groups. For example, if "p" and "t" are in the same phone group (plosives), the words "top" and "pot" have all phone group/positions in common, and the predicted discriminability is very low.

For some languages pronunciation dictionaries are publicly available. However, the complete set of command words in a service or product is limited, and the individual pronunciation of the command words can be found by consulting an expert phonetician. This person can also help in specifying groups of phones that can be considered "very similar".

3.3.3 *Applying the acoustic discrimination*

The discriminability measure can be used to find the optimally performing command words for each recognition context, which is a complex procedure. An example can help to clarify the procedural difficulties. Suppose, for instance, that in a media browsing application the functions "move to first message" and "exit application" are available simultaneously. Suppose further that for the first function the commands "top" and "first" come out of the confidence test with preferences 70 % and 30 %, while for the second function the commands "quit" and "stop" appear to have preference levels 25 % and 75 %, respectively. Without paying attention to acoustic discrimination, the command words "top" and "stop" would be the preferred ones. If the acoustic discriminability is taken into account, however, which word is going to be replaced by an alternative command with lower subjective preference? The percentages for the alternative words, 30 % and 25 % respectively, appear very similar, and moreover they may not be the only important factor. There might be other words for which discriminability plays a role, for example, a command word "quick" with high subjective preference. This means that discriminability optimization is a process that should be applied to the whole menu structure, possibly involving different contexts and even different applications.

It is quite difficult to formally define a procedure for optimizing the command vocabulary words, because many more factors should then be incorporated such as frequency of occurrence of the commands and likelihood that other applications will be available. A more pragmatic approach to the problem therefore is the following procedure:

- a) For each context, start with the command words suggested by the confidence rating test.
- b) Find possible pairs of commands that give rise to acoustic discriminability problems.
- c) Choose an alternative for one of the command words, with minimum repercussion with respect confidence rating.
- d) Repeat step b) and verify that there are no other commands that clash acoustically with the alternative command word.
- e) Repeat step a) to verify that all functions that have new alternative commands do not occur in other contexts or have no acoustic discriminability problems there.

This procedure assumes that there is a relatively low probability that two command words will have low acoustic discriminability.

4. Observations

In the development of the previous standard for five languages, the Storyboard method was discarded in favour of carefully worded descriptions. For 28 languages, it may be too time-consuming to use carefully worded descriptions, and the Storyboard method may be reconsidered since one set of drawings can be applied to multiple languages. This choice is still under consideration.

Within the available time-frame for this project, we are limited with respect to our data collection and the choice of methodology for defining the vocabularies to be included in this standard. It is expected that the number of subjects for some of the targeted languages will be limited, as access to a large pool of native speakers from all 28 countries may not be achievable within our timeframe. Furthermore, it is unlikely that, in the test phase of the research, we will have comparable state-of-the-art ASR systems available for each or the targeted languages. In consequence, we will most likely follow the *pronunciation dictionary test* in assessing the phonetic discriminability between words within a language.

However, despite limitations on resources, we are confident that the standard will be produced, that it will contribute to the development of voice user interfaces and that it will improve accessibility and usability for the user.

5. Invitation to participate

Since the development of the current standard, technology has progressed and voice user interfaces are applied to more applications, in an increasingly embedded fashion. We intend to review the vocabularies from the previous standards and possibly to change or extend them. We therefore invite active participation from all interested parties in the development of this standard. It is an open process and contributions can be made to the working group's leader, Mike Tate (mtate@essex.ac.uk).

6. Acknowledgments

We are grateful to the European Commission, ETSI, and our employers for supporting this work and enabling our participation. We would like to specifically acknowledge the substantial contribution of the developers of the ES 202 076 standard, David van Leeuwen, Catriona McDermid, Bruno van Niman, Lutz Groh, Wally Mellors and Scott McGlashan for their work on the methodology for this type of project.

References

- [1] ETSI ES 202 076: "Human Factors; User Interfaces; Generic spoken command vocabulary for ICT devices and services" (version 1.1.2) 2002.
- [2] ETSI EG 202 132: "Human Factors; User Interfaces; Guidelines for generic user interface elements for mobile terminals and services".
- [3] McDermid, C. and M. Goldstein: "The 'storyboard' method: Establishing an unbiased vocabulary for keyword and voice command applications" HCI Industry Day and Adjunct Proceedings, pp 104 – 109, 1996
- [4] Guzman, S. J. Warren, R., Ahlenius, M., and Neves, D: "Determining a set of acoustically discriminable, intuitive command words". Proceedings AVIOS, pp 242 – 250, 2001.
- [5] EC Communication on e-Accessibility, F5/PB/pla D(2004) 535262
- [6] ETSI EG 202 191: "Human Factors (HF); Multimodal interaction, communication and navigation guidelines"
- [7] Directive 2002/22/EC of the European Parliament and of the Council on universal service and users' rights relating to electronic communications networks and services (Universal Service Directive), Journal of the European Communities, Official Journal L 108 , 24/04/2002 P. 0051 - 0077, March 2002.