VOIce-based Community-cEntric mobile Services for social development

Grant Agreement Number 269954

Deliverable No D3.4
Report on user acceptance studies – cycle 2

May 2013

SEVENTH FRAMEWORK PROGRAMME

THEME ICT-2009.9.1 – International cooperation

PROJECT DELIVERABLE REPORT

Project

Grant Agreement number	*2699542*
Project acronym:	*VOICES*
Project title:	*VOIce-based Community-cEntric mobile Services for social development*
Funding Scheme:	*Collaborative Project*
Date of latest version of Annex I against which the assessment will be made:	*18 February 2011*

Document

Deliverable number:	D3.4
Deliverable title	Report on user acceptance studies – cycle 2
Contractual Date of Delivery:	April 2013
Actual Date of Delivery:	27 May 2013
Editor (s):	
Author (s):	Etienne Barnard
Reviewer (s):	Jenny de Boer
Work package no.:	WP3
Work package title:	Speech Technologies
Work package leader:	NWU
Work package participants:	W3C, FT, CSIR, PTIN
Distribution:	PU
Version/Revision:	0
Draft/Final:	Final
Total number of pages (including cover):	14
Keywords:	Automatic speech recognition, Tabale, usability, acceptance studies

CHANGE LOG

Reason for change	Revision	Author	Date
Document creation	0.1	Etienne Barnard	10-05-13
Update recognition statistics	0.2	Etienne Barnard	12-05-13
Release version, incorporating editors' suggestions	1.0	Etienne Barnard	27-05-13

DISCLAIMER

This document contains description of the VOICES project work and findings.

The authors of this document have taken any and all available measures in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the VOICES consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

VOICES is a project funded in part by the European Union.

TABLE OF CONTENTS

SUMMARY

This report presents the results of the first round of user acceptance studies that were performed in Bamako and Tominian, Mali during April and May 2013. With the user acceptance studies in WP3 we wish to assess the usability and desirability of speech technology as evaluated by the target populations for the VOICES pilot studies. This second cycle evaluated these aspects of our speech-activated meeting scheduling application, *Tabale*; in addition, we also estimated its speech-recognition accuracy on a limited sample set.

To this end, twelve subjects were recruited in Bamako and Tominian, asked to interact with *Tabale* according to a specified protocol, and then given a questionnaire to complete. The questionnaire investigated both the usability of the application and the respondents' opinion on its practicality. Overall, respondents found the system to be very usable; this was confirmed by the large number of practical use cases suggested by the respondents and their successful completion of tasks specified by the protocol.

The accuracy of the Bambara speech-recognition component was assessed by comparing the recognition results produced by the automatic recognizer with manual transcriptions of the recorded utterances. The baseline recognizer (developed as one of the components of deliverable D3.3) was found to be approximately 90% accurate on the small sample set that was collected; this recognition accuracy could be improved further by adapting to the samples that were collected in a speaker-independent fashion.

INTRODUCTION

In our report on the first cycle of usability studies in WP3 (VOICES Deliverable 3.2), we summarized why it is important to involve end users in evaluating how well speech-technology components function – especially in a project such as VOICES, which involves a user population without prior exposure to such technology. The current cycle continued our work in this regard – in particular, we investigated the usability of our speech-activated meeting scheduling application, *Tabale*. (This application is described in more detail below.) The investigation had two major goals:

- To investigate whether our target user population in Mali can successfully use *Tabale* – in particular, the automatic speech recognition (ASR) component that is used to capture user responses.
- To measure the accuracy that is achieved when using the Bambara-language ASR system in this real-world task.

This study was planned for January 2013; however, various technical issues and logistical challenges in Mali caused its postponement to April and May 2013. All tests were conducted by staff members of the Malian NGO Sahel Eco, on location in Bamako and Tominian, Mali.

The next chapter describes the method that was used in order to meet the goals described above; we then present the results that were obtained during our studies. The concluding section summarizes the main findings of these analyses, and details the recommendations for future use of this technology and of the *Tabale* service.

METHOD

The *Tabale* application was designed and implemented by the WP5 team, to assist with the arrangement of meetings. It consists of two interacting software modules: the first is a Web interface, aimed at meeting organizers (who are assumed to have Internet access) and the other a telephone interface for invitees, who do not typically use the Internet. Our study is restricted to the second component, since the first does not involve speech technology. The telephone interface functions as an Interactive Voice Response (IVR) system, which performs the following functions:

- All invitees selected by the meeting organizer are called in sequence.
- When an invitee answers the telephone, a pre-recorded message from the organizer, announcing the meeting, is played to the invitee.
- The invitee is then asked whether (s)he is planning on attending the meeting, and requested to respond with a "yes", "no" or "maybe".
- Finally, the invitee is given the option to leave additional comments, if desired.

The entire interaction is conducted in the regional language, Bambara.

Staff members of Sahel Eco conducted the usability tests in Bamako and Tominian, Mali. In each location, three females and three males were enrolled to participate in the test; each of the twelve respondents were called three times by the system, and requested to respectively respond as if they *would attend*, *would not attend*, and were *unsure about attendance* on the three calls. After completion of the tests, each respondent completed a brief questionnaire (see Appendix A). All responses were tallied, and are summarized below; we also measured the accuracy of the ASR system in two different configurations, and report on those results as well. One questionnaire was lost prior to data processing – hence, our results contain 11 responses to each question.

In addition, the accuracy of the speech recognition (performed by the Bambara ASR system that was one of the components of Deliverable D3.3) was assessed by manually transcribing each utterance, and comparing each recognized response with the corresponding manual transcription.

RESULTS AND DISCUSSION

Responses to questionnaires

Figures 1 to 3 below summarize the objective results that were obtained during our usability test. Figure 1 shows histograms of the responses we obtained to the question "Q1: Was it easy to follow the **instructions** of the system?" In this and subsequent figures, a response of "1" means "definitely yes", "2" means "yes", "3" is "neutral", "4" is "no" and "5" corresponds to "definitely no". Thus, 7 out of the 11 respondents found the instructions very easy to understand, and only 1 one of the respondents was neutral on this question. (This respondent, as well as one other respondent, commented on the fact that the system was not clear on the fact that they should speak *after the beep.*)

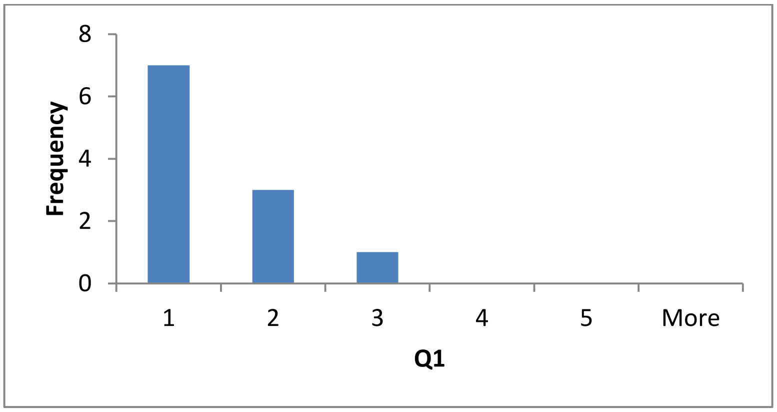

Figure 1: *Responses to the question related to simplicity of instructions. A response of "1" means "definitely yes", "2" means "yes", "3" is "neutral", "4" is "no" and "5" corresponds to "definitely no".*

Similarly, Figure 2 summarizes the responses to the question "Q2: Was the **pace** of conversation OK?" Again, all respondents found the pace to be satisfactory, with five respondents selecting "Definitely yes" and the remaining six selecting "Yes". However, two respondents did comment that they found the pace of the interaction to be rather quick, whereas another commented that it was on the slow side.

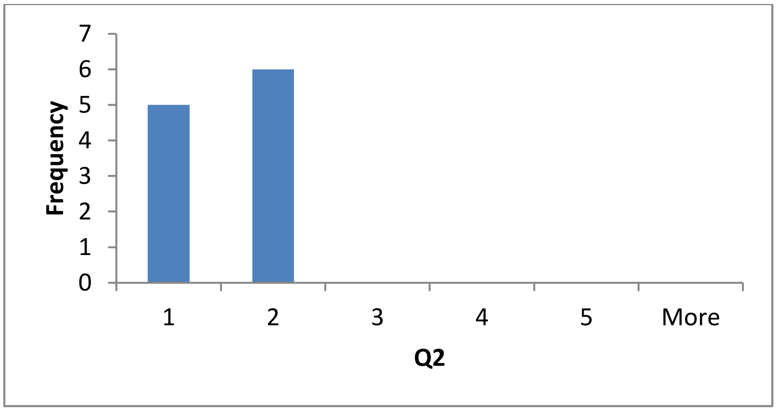

Figure 2: *Responses to the question on the pace of the interaction. A response of "1" means "definitely yes", "2" means "yes", "3" is "neutral", "4" is "no" and "5" corresponds to "definitely no".*

Results for the question "Do you think that this service is **useful**?**"** are summarized in Figure 3. Again, responses are quite positive.

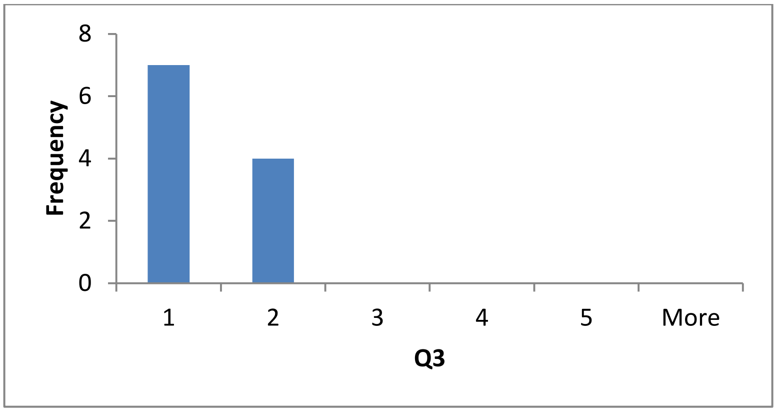

Figure 3: *Responses to the question related to speech rate, for recorded speech (top panel) and generated speech (bottom panel). A response of "1" means "definitely yes", "2" means "yes", "3" is "neutral", "4" is "no" and "5" corresponds to "definitely no".*

This impression was confirmed by responses to the two open-ended questions that were asked at the end of the questionnaire ("Q4: If such a service were available to you, how would you use it" and "Q5: Do you have any other comments?"). All 11 respondents were able to think of scenarios in which the service would be useful to them or their organizations. For example, one respondent suggested that the system could be used "To coordinate work on our building sites" and another responded that "Our Group Enterprise could use it to orgnaise meetings, training sessions for members and information for clients". Six of the respondents had positive comments on the service (remarking, for example, on its speed and economic advantages). The only two negative comments in response to Q5 related to technical problems: one caller apparently had problems with GSM reception, whereas the recording for the other was not performed successfully.

Recognition accuracy results

Due to various technical challenges, we were not able to retrieve all the recordings that were made during the usability tests (some recordings were apparently lost due to a power failure, and others may have been lost in GSM network problems); our analysis is therefore based on a mixture of recordings made during the usability tests and recordings by Sahel Eco staff members at other times. As a result, we had access to a set of 69 recordings from 10 different speakers (5 females and 5 males). Of these recordings, 50 contained speech and the other 19 were empty (presumably because of user-interface issues, users testing the system, problems with mobile-telephone reception, etc.); our analysis below concentrates on the recordings that did contain speech.

Our first observation is that all these recordings were *in vocabulary* – that is, the users spoke one of the expected words, and nothing else. The IVR prompt explicitly requested callers to respond with "yes, no or maybe" (in Bambara), and all callers abided by that request. Our baseline system was able to perform fairly accurate recognition on these utterances: it successfully recognized 45 out of the 50 utterances, corresponding to a recognition accuracy of 90%. The confusion matrix produced by the system is contained in Table 1 below – we see that "yes" ("awo") and "maybe" ("m'a don") were both recognized as "no" ("ayi") once, and that two occurrences of "m'a don" and one of "ayi" were incorrectly recognized as "awo".

		Recognizer Output			
		awo	ayi	m'a don	No output
True utterance	awo	19	1	0	0
	ayi	1	15	0	0
	m'a don	2	1	11	0

Table 1: *Confusion matrix produced by baseline ASR system*

We also experimented with Maximum A Posteriori (MAP) adaptation of the recognizer, using a cross-validation protocol. (MAP adaptation is a common algorithm in speech processing – it allows an existing system to be refined using a small set of task-specific data.) Thus, the data from each speaker is used as test data in turn, whereas the data from the remaining speakers is used to adapt the acoustic models. Even though the adaptation data is extremely limited, adaptation did produce an apparent improvement in the system – now, only one utterance was not successfully recognized, as shown in Table 2.

		Recognizer Output			
		awo	ayi	m'a don	No output
True utterance	awo	20	0	0	1
	ayi	0	16	0	0
	m'a don	0	0	13	0

Table 2: *Confusion matrix produced by MAP-adapted ASR system*

CONCLUSION

Our usability tests have confirmed that the targeted user population find the speech user interface in *Tabale* intuitive and easy to use. The command to speak was sufficiently clear to elicit expected phrasings in all cases for which speech was recorded. The users' suggestion that an explicit request to "speak after the beep" be included should, however, be considered – especially in view of the large number of empty recordings that were saved by the system.

The ASR recognition results obtained by the system were very encouraging. Although the vocabulary of the system is very small (only three words), the words are somewhat confusable, and the acoustic models were developed with absolutely no bias towards these words. This bodes well for the future applications of such a recognizer, in which larger but less confusable vocabularies are likely to occur.

The positive outcome observed from MAP adaptation to match the acoustic models with the vocabulary employed was also interesting. This outcome suggests that such adaptation may play an important role in resource-constrained environments: if it is possible to collect even a small amount of data with "generic" acoustic models, adaptation may be sufficient to develop significantly improved models. Of course, our data set was very small, and additional tests will be required to verify the capabilities of such an approach.

The most important overall finding of our usability studies is that speech technology can indeed play a useful role in resource-constrained environments, since users were able to use the speech-enabled systems successfully and were positive about the demonstrated use cases. Along the way, we have learned several lessons, and come across a number of surprises; the potential of speech technology was nevertheless firmly established during the course of our work.

ACKNOWLEDGMENT

It is a great pleasure to thank Mary Allen and her staff at Sahel Eco – they were able to overcome challenges with development process and the various platforms used in very trying circumstances, and repeatedly went the extra mile with the hard parts of usability testing.

APPENDIX A: QUESTIONNAIRE

VOICES Tabale Usability Questionnaire

Respondent: **Observer:**

Date & time:

Key: 1-Definitely YES 2-Yes 3-Neutral 4-No 5 Definitely NO

No	Question	1	2	3	4	5	COMMENTS
1	Was it easy to follow the **instructions** of the system?						
2	Was the **pace** of conversation OK?						
3	Do you think that this service is **useful**?						
4	If such a service were available to you, how would you use it						
5	Do you have any other comments?						