VOIce-based Community-cEntric mobile Services for social development

Grant Agreement Number 269954

Deliverable No D3.5
Proposal documents to standards committees

June 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:	D3.5
Deliverable title	Proposal documents to standards committees
Contractual Date of Delivery:	June 2013
Actual Date of Delivery:	June 2013
Editor (s):	
Author (s):	Etienne Barnard
Reviewer (s):	Filipe Cabral Pinto
Work package no.:	WP3
Work package title:	Speech Technologies
Work package leader:	NWU
Work package participants:	W3C, FT, CSIR, PTIN
Distribution:	PU
Version/Revision:	1.0
Draft/Final:	Final
Total number of pages (including cover):	20
Keywords:	Speech technologies, under-resourced languages, standards and best practices

CHANGE LOG

Reason for change	Revision	Author	Date
Document creation	0.1	Etienne Barnard	8-06-13
Updated version, incorporating suggestions from S Boyera	0.2	Etienne Barnard	10-06-13
Final review	0.3	Filipe Cabral Pinto	12-06-13
Release version, incorporating reviewer's suggestions	1.0	Etienne Barnard	14-06-13

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

CHANGE LOG 3
DISCLAIMER 4
TABLE OF CONTENTS 5
SUMMARY 6
INTRODUCTION 7
THE CHALLENGES OF SPEECH TECHNOLOGY DEVELOPMENT FOR UNDER-RESOURCED LANGUAGES8
THE STANDARDS-BASED APPROACH IN VOICES 10
Relevant standards for voice services 10
TOOLS AND BEST PRACTICES FOR UNDER-RESOURCED SPEECH TECHNOLOGY DEVELOPMENT 12
Developing small-vocabulary ASR systems in under-resourced languages 13
Developing slot-and-filler TTS systems in under-resourced languages 15
RESULTS: USABILITY AND ACCURACY 17
Usability of slot-and-filler TTS in Bambara, Bomu and African French 17
Usability and accuracy of small-vocabulary ASR in Bambara 17
CONCLUSION 19
REFERENCES 20

SUMMARY

The VOICES project has demonstrated the potential of speech technology in the developing world, and has also delivered a number of technology components that were found to be useable in field trials. Based on these experiences, we are confident that existing standards for voice technologies (including VoiceXML, SSML, PLS and related W3C standards) are sufficiently flexible to be of great value for under-resourced languages. Therefore, the contents of the deliverable have been modified from the originally planned "proposals to standard committees" to be more wide-ranging proposals on how speech technology should be developed for under-resourced languages.

Several of the open-source tools that were in existence at the start of the project (such as *HTK* and *Festival*) were found to be useful during our developments; we have also added some additional tools that should help others with similar goals in future. We provide brief guides on how small-vocabulary Automatic Speech Recognition systems and slot-and-filler Text-to-Speech systems can be developed using these tools.

The languages that were the focus of VOICES could be classified as moderately under-resourced; for severely under-resourced languages, in which textual resources are very hard to come by, could probably not be addressed with the tools that we have utilized. Continued research on speech technology creation for such languages should remain a topic of focused research, along with expanded development and application of voice technologies in languages similar to those studied during VOICES. Finally, we recommend a systematic push towards speech technology development in Africa, where languages occur in families with strong relationships. By starting with regional languages such as Wolof and Swahili, it should be possible to systematically cover many of the large languages reasonably quickly, thereby giving crucial tools to prospective developers of voice services.

INTRODUCTION

Access to reliable, up-to-date information is a significant challenge for most people in the developing world, where modern information technology is not widely available. Thus, people often travel long distances to obtain information about critical matters such as health care, financial assistance or market. Voice-based services have the potential to become a primary source of such information in the developing-world context. Although traditional computer infrastructure is often scarce in emerging countries, telephone networks (especially cellular networks) are spreading rapidly, thus creating tremendous opportunities for voice-based information access.

The VOICES project was one of the first large-scale projects aimed at realizing this potential; as such, it provided an ideal opportunity to learn how the tools and standards for speech technology can be applied to the under-resourced languages that are most common in emerging regions. The current document summarizes the challenges that we faced, the tools and processes we developed to address those challenges, and our main findings. In the following Section, we highlight some of the challenges that are faced when developing speech technology for under-resourced languages. We then discuss the approach that was taken for speech technology development in VOICES relevant open standards. The section thereafter discusses the tools and practices that we have developed for the creation of speech technology for under-resourced languages. We then summarize the results that were obtained using these tools, both in applications and in off-line evaluations. Finally, the conclusion contextualizes this work with respect to the entire universe of spoken languages, and makes recommendations for further research and development, with a specific focus on voice technology in Africa.

The most important speech technologies for information access are Text-to-Speech (TTS) synthesis and Automatic Speech Recognition (ASR); these two technologies are therefore the focus of the current report. The objective of this deliverable is to promote the results of VOICES within the speech community; since we did not find any significant need for changes to the relevant standards, the contents of the deliverable have been modified from the originally planned “proposals to standard committees” to be more wide-ranging proposals on how speech technology should be developed for under-resourced languages.

THE CHALLENGES OF SPEECH TECHNOLOGY DEVELOPMENT FOR UNDER-RESOURCED LANGUAGES

Although speech technology is now widely used in the developed world, it is still mostly unavailable in the under-resourced languages of the developing world. The primary cause for this discrepancy is the technical complexity of speech-technology development, which results in a number of challenges when such development is aimed at under-resourced languages. The most important challenges include the following:

- The codification of appropriate linguistic knowledge - especially phonological and phonetic information - which will often require original research for the languages of the developing world.
- The collection and development of basic resources such as word lists, phone sets, pronunciation dictionaries and corpora for resource-scarce languages. Whereas electronic repositories of such information are available for the well-resourced languages (and are often in the public domain), this is not the case for most of the languages of the world. In some cases, the relevant information can be gleaned from paper texts, or from relevant proprietary sources. However, for many under-resourced languages, these resources will have to be created from scratch.
- The development of speaker-independent automatic speech recognition (ASR) systems that function reliably in the local languages of the developing world. As we discussed in Deliverable 3.1, language-independent tools, as created for well-resourced languages, are typically suitable for the core ASR engine. However, three language-specific components are required to enable ASR in a chosen language, namely the acoustic model, language model and pronunciation dictionary (see Fig. 1). The creation of these components requires the basic resources described above as well as relevant tools and expertise to extract or calculate the desired models.

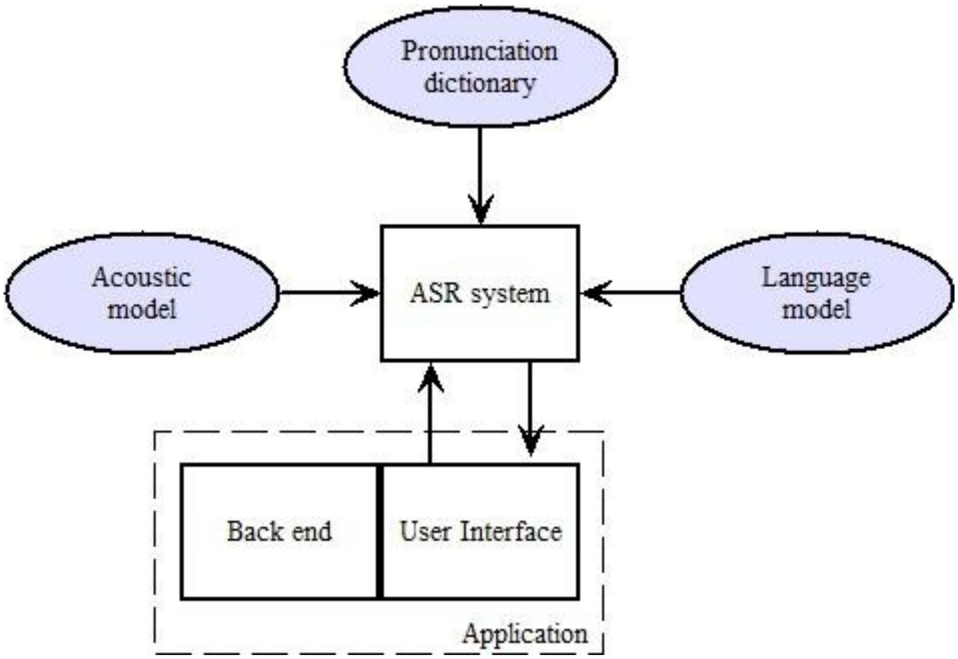

Figure 1: Block diagram of a speech-enabled application; the most important language-specific components are shaded. (From Deliverable 3.1)

- The development of text-to-speech (TTS) systems that are easily understood. As with ASR, language-specific models are used in conjunction with language-independent algorithms to create such systems. These language-specific components again include pronunciation and acoustic models, although the latter are somewhat different in TTS compared to the ASR models. In addition, text processing is an important aspect of the TTS processing chain.

In the developing world, these challenges must generally be faced in environments where appropriate skills are scarce or non-existent, and where material resources are limited. Therefore, a consistent theme in speech research for the developing world is the development of assistive tools and the automation of processes wherever possible. Another important requirement for the development of under-resourced languages is cost efficiency: hundreds of millions of dollars were invested in the creation of the initial speech-technology systems in languages such as English, Japanese and German. Given that there are hundreds or even thousands of languages in which speech technology is required, in environments with severe budgetary constraints, it is clear that more economical options are required.

THE STANDARDS-BASED APPROACH IN VOICES

To meet the challenges described in the preceding section, we decided to use open-source tools and standards-based methods as far as possible. As we review below, numerous relevant standards have been created in the past two decades to support the development of voice services in the developed world; the question, therefore, was whether these standards would also be suitable for the languages and environments that are pertinent to the VOICES project. In particular, the languages covered by our use cases are Bomu and Bambara (also known as Bamanankan) – these are Niger-Congo languages. Bambara, which functions as a regional language in Mali and surrounding countries, has significantly more resources than Bomu (which has fewer than 500,000 native speakers according to Ethnologue); however, both languages qualify as under-resourced languages, with no content relevant to speech technology available prior to the current project.

Relevant standards for voice services

There are two classes of standards that are of principal relevance to voice services: (1) those that relate to the application and its software / hardware environment and (2) those that are relevant to the language technologies used within the application. The most important standards in the first class are the following:

- **VoiceXML**, which is used to specify interactive voice dialogues. An XML-based syntax and design philosophy enable developers to use tools and architectures that are widely used for Web services to also support voice services.
- **Call Control eXtensible Markup Language (CCXML)** is used to specify the call control surrounding an application. This control includes standard activities such as call answering and disconnection, but also more sophisticated capabilities such as call transfer.

The standards that are commonly used in implementing voice technologies are all W3C Recommendations; they include:

- **Pronunciation Lexicon Specification (PLS),** which allows developers to specify how words are pronounced, with both the alphabet used for word orthographies and the phoneme sets used for pronunciation description under control of the programmer. These pronunciations can be used in both speech synthesis and speech recognition.
- **Speech Synthesis Markup Language (SSML)** is used to control aspects of synthesized speech such as volume, pitch and rate; it therefore allows developers to control how the requested text is spoken at a low level, and with a fine grain.
- **Speech Recognition Grammar Specification (SRGS)** is the common standard for specifying the utterances that can be recognized by a speech-recognition system – that is, the sequences of words that are expected by the recognizer.
- **Semantic Interpretation for Speech Recognition (SISR)**, which is generally used within SRGS to extract the meaning from recognized utterances.

Although these standards were all developed within the context of well-resourced languages, their designers were well aware of the requirements of internationalization; in each case that we investigated in detail, these requirements were met in ways that support under-resourced languages as well. In particular, universal support for UTF-8 implies that character sets from smaller languages are handled without any difficulties. Also, the layered design of these standards supports different language classes well. For example, in speech recognition, the hierarchy PLS -> SRGS -> SISR is equally suitable for writing styles that are more or less phonetic, for languages with radically different word orders, and for other forms of variability. Finally, the low level of control afforded by SSML allows for a wide range of different linguistic constructs to be implemented.

TOOLS AND BEST PRACTICES FOR UNDER-RESOURCED SPEECH TECHNOLOGY DEVELOPMENT

As described in Deliverable 3.1, the development of speech technology requires the collection of several resources. The list of resources required typically includes some or all of the following:

1. Grapheme and phoneme sets;
2. Text corpora;
3. Word lists;
4. Pronunciation lexicons (dictionaries) or rules;
5. Transcribed speech corpora.

Our experience during VOICES suggests that (1), (3) and (4) in the list above are relatively easy to obtain, even for under-resourced languages. In particular, sources such as introductory grammars, language-learning courses and even Wikipedia provide generally reliable information, which can be cast into the desired format for further processing with standard programming tools. (Obviously, there is also a class of extremely under-resourced languages for which even such basic resources are not available – we return to this matter in the Conclusion below.)

For the creation of text corpora, our hope was that the crawling of Web resources would provide sufficient raw materials to serve as starting point. However, even for a widely spoken language such as Bambara, this did not turn out to be true, and the scanning of paper documents and a significant amount of manual curation were required to develop suitable text corpora for technology development. Although this situation is likely to change reasonably quickly as the Web continues to penetrate ever deeper into societies worldwide, the Web is currently not a sufficient source for text corpora in many under-resourced languages. Hence, it is advisable that developers of speech technology should cultivate partnerships with local governments, publishers and other likely sources of large amounts of electronic text in the targeted language(s).

Speech corpora are, of course, even less common in under-resourced languages. Fortunately, a number of open-source tools have been created in recent years to assist with the creation of such corpora. The *VOCS* platform, which was developed during the current project, is suitable for environments in which illiterate respondents or respondents without internet access are to contribute the speech samples. A tool such as *Woefzela* is even more convenient to use for the collection of speech corpora, but requires literate respondents and mobile internet access (since respondents are required to *read* the prescribed prompts, and smartphones are used during the data-collection process).

Once these resources have been created, the development of text-to-speech (TTS) or automatic speech recognition (ASR) systems can commence. Again, a wide range of useful open-source tools and toolkits have been released in recent years; several of those tools are discussed in Deliverable 3.1, and the recommendations in that report remain valid. In addition, a number of more recent developments should also be noted by those who wish to develop speech technologies for under-resourced languages:

- The value of *HTK* was confirmed during the current project; it was critical to the development of our Bambara ASR system. However, we also found that the creation of an end-to-end ASR system requires a fair amount of additional software; we have therefore released a set of scripts called *ASR-template*, which wrap around *HTK* and simplify the development process considerably.
- The past two years have seen much activity around the development of *Kaldi*, a new open-source toolkit for ASR development. Although *Kaldi* contains some more modern algorithms which are not incorporated in *HTK* (such as support for Finite-State Transducers and Deep Belief Networks), it currently is also notably less mature (and thus, stable) than *HTK*. Our expectation is that *HTK* will remain the tool of choice for under-resourced ASR development in the near future, but that *Kaldi* will supersede it as the de-facto standard in time to come.
- For TTS, the *Festival Speech Synthesis System* (including *Festvox* for building voices) remains the most popular open-source toolkit. However, during the current project we have found that there is a significant need for a simpler set of tools that can be used to build small, domain specific slot-and-filler TTS systems. We have consequently developed and released a set of tools that are suitable for this task; using these tools, it is possible for a software developer with limited knowledge of language technology to develop such a slot-and-filler TTS system in any language.

Given the availability of these tools, we next summarize the steps that we recommend for the development of small-vocabulary ASR systems and domain-specific TTS systems that proved to be so useful in VOICES.

Developing small-vocabulary ASR systems in under-resourced languages

The steps that are required to develop a sub-word based ASR system are summarized in Fig. 2. The external inputs to the process are shown as coloured ovals, and the way these are obtained will depend strongly on the language in question, as described above. For steps (1) and (3) in Fig. 2, bespoke software is required to select text that is suitable for the application of interest, but the other steps can all benefit from open-source tools: DictionaryMaker for lexicon creation (step 2), the VOCS platform for recording speech (step 3) and HTK with ASR-template for acoustic modeling. The grammars used for small-vocabulary systems are usually short lists of allowable words; hence, the language-modeling step is trivial in this case: the expected responses are listed in a text file, and compiled into a recognition network with HTK tools. Of course, some amount of iteration should be expected in this process: during field trials, it may turn out that users frequently employ phrasings that are not included in the list of expected responses. In that case, the additional phrasings can simply be added to the list, or the prompting can be changed to elicit the expected responses more reliably.

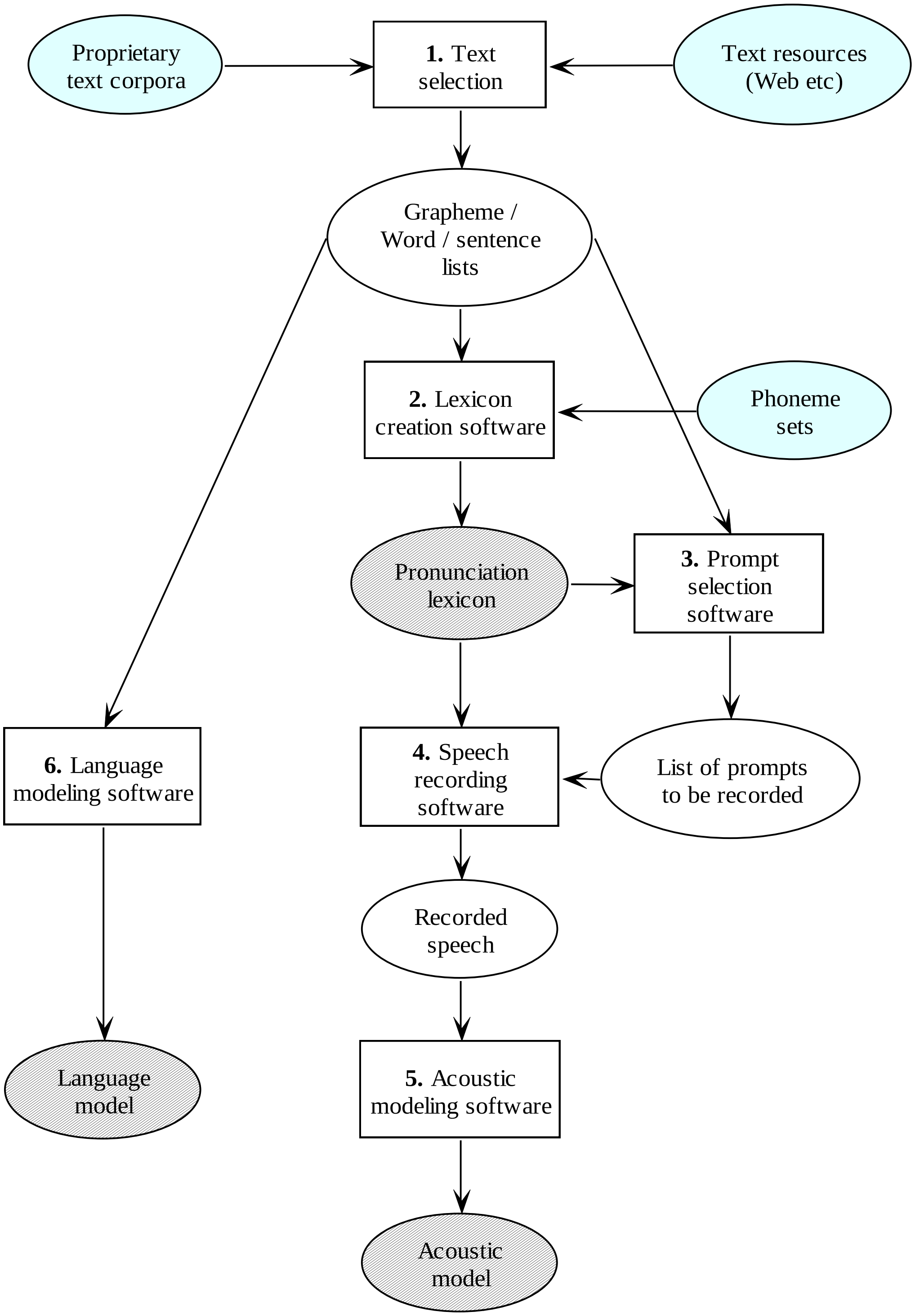
Proprietary text corpora
1. Text selection
Text resources (Web etc)
Grapheme / Word / sentence lists
2. Lexicon creation software
Phoneme sets
Pronunciation lexicon
3. Prompt selection software
6. Language modeling software
4. Speech recording software
List of prompts to be recorded
Recorded speech
Language model
5. Acoustic modeling software
Acoustic model

Figure 2: Recommended development process for the creation of an ASR system in an under-resourced language

Developing slot-and-filler TTS systems in under-resourced languages

During the VOICES project we found that slot-and-filler TTS systems are useful for the types of voice services considered. Such systems are considerably simpler to develop than full-fledged subword-based systems, as can be seen from the process diagram in Figure 3.

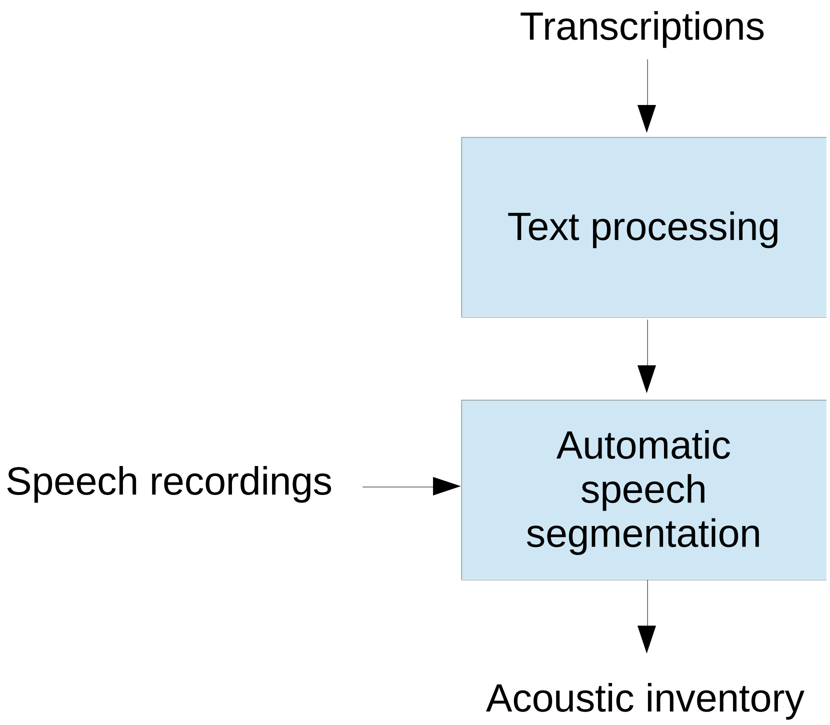

Figure 3: Recommended development process for the creation of a slot-and-filler TTS system in an under-resourced language (from Deliverable 3.2)

In this case, the transcriptions are derived directly from the application specification – thus, there is no need for a text corpus, which simplifies the development process considerably. Also, the speech recordings are made by a single speaker, and can thus be done on a suitable personal computer and standard recording software.

The text-processing and speech-segmentation steps are managed with the S&F TTS tools that were created during the current project. These tools interface with HTK, on the one hand, and a formatted text file, on the other, to create an inventory of acoustic items that are subsequently used during synthesis, as shown in Fig. 4. As in Fig. 3, the processing steps in the coloured blocks are implemented by the S&F TTS toolkit.

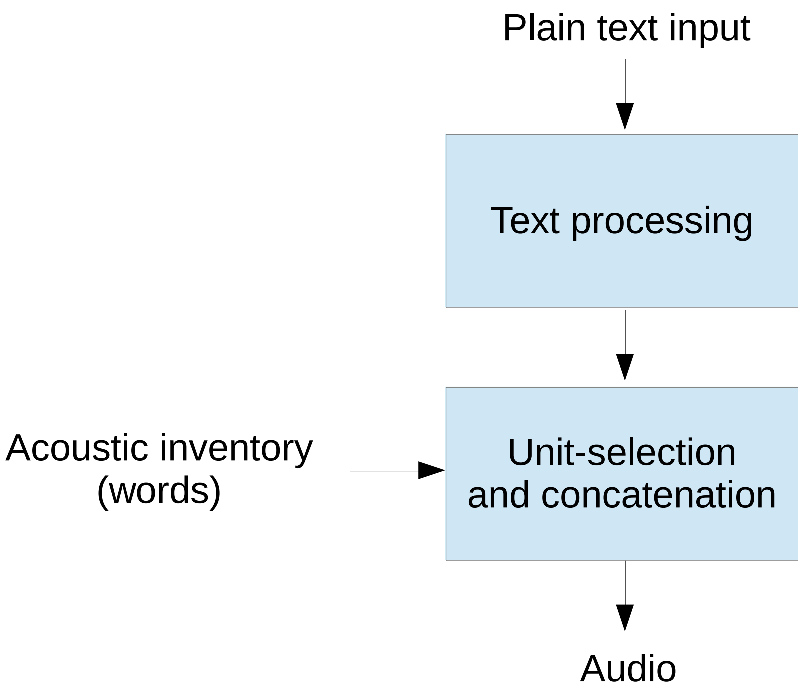

Figure 4: Synthesis process with a slot-and-filler TTS system (from Deliverable 3.2)

RESULTS: USABILITY AND ACCURACY

To evaluate the performance of the speech-technology tools (voice packs) that were developed using the approach described above, a series of usability tests were undertaken. The full results of those tests are contained in Deliverables 3.3 and 3.4; here, we summarize the main findings to indicate the capabilities that can be achieved in this way.

Usability of slot-and-filler TTS in Bambara, Bomu and African French

Our first usability studies investigated three slot-and-filler TTS systems, which were used in a market-price application. These systems take as input a spreadsheet containing various pieces of information related to agricultural products that are available for sale – including the location of production, the quantity, quality and price of the available produce - and the contact information of the person selling the produce. Using a slot-and-filler approach, an audible communiqué is then created in three languages: Bambara, Bomu and African French.

The usability test involved respondents in the target population who frequently speak these languages; these respondents were asked to comment on the intelligibility, naturalness and speed of the synthesized voices, and also to compare the natural and synthesized versions of the same bulletin. For all three languages, respondents expressed a high degree of satisfaction with the synthesized voices: all respondents found the synthesized communiqués to be either very intelligible or acceptably intelligible, and most respondents also found them to be acceptable in terms of naturalness and speed.

Of course, these TTS systems are not as natural as the human voice, and therefore respondents generally preferred naturally recorded communiqués to synthesized versions when given the choice. However, the high level of acceptability of the synthesized communiqués, along with the great convenience of directly synthesizing versions in all three languages from a single input document, demonstrates that such TTS systems have great potential.

It should be noted that our synthesis was based on recordings of radio announcers who were known in the regions where the tests were undertaken. Some of the respondents recognized these voices, and commented positively on this fact; the convenience of using such familiar voice artists is a major benefit of the simplicity of slot-and-filler TTS systems.

Usability and accuracy of small-vocabulary ASR in Bambara

The accuracy of an ASR system can be assessed on-line, in live usability tests, and off-line, using previously recorded data. We performed both types of evaluation on the small-vocabulary ASR system that was developed during the project.

On the accuracy front, we experimented with three tasks: two artificial ten-word tasks that were designed to investigate the small-vocabulary accuracy that can be expected, and a three-word task that was required for the *Tabale* application. For the ten-word tasks, we employed a test set extracted from the data that had been collected during our ASR development process. We measured 95.6% and 89.3% accuracy on these two tasks, respectively. The *Tabale* recognition task was evaluated on data collected during usage of the application, and we achieved 90% recognition accuracy with no adaptation to the task data, and 98% recognition accuracy after Maximum A Posteriori (MAP) adaptation of the recognizer.

Finally, the usability assessment of the *Tabale* application demonstrated that most of the target users were easily able to use the speech-technology enable application. All respondents were able to use the system successfully (although a few had to experiment with the system a little before achieving success), and ten out of the eleven respondents in the test responded either "yes" or "definitely yes" when asked whether the system was simple to use.

CONCLUSION

The successful creation of ASR and TTS systems during the current project and their deployment in real-world use cases demonstrate that speech technologies for under-resourced languages are both achievable and useful. We have also found that the relevant standards were quite suitable for our developments, and that many of the existing software tools could be adapted to these languages.

Of course, our attention was limited to only two under-resourced languages (Bambara and Bomu) – hence, one should question how well our conclusions will generalize to other languages in this category. Two classes of languages need to be considered:

- Languages with a similar degree of resourcing as those we considered, but vastly different linguistic structures (e.g. non-alphabetic writing styles or other types of phonetic distinctions), are likely to be handled as successfully as those that we considered. In this regard, the widespread acceptance of UTF-8 is an important advance.
- Languages with significantly fewer resources (including languages that do not even have a generally-accepted writing system, languages with small speaking populations, and languages in which the vast majority of speakers are illiterate), on the other hand, are probably still beyond the reach of current speech technology. The reliance on substantial speech corpora to create other resources such as lexicons and prompts is especially problematic for such languages, and it is probably premature to design tools and standards for them.

In conclusion, then, we believe that a strong basis exists for the development of speech technology in under-resourced languages, and the current project has shown that such development could be of great practical use. While it is imperative that research continues to push on the lower limits of resources required for technology creation, there is a wide band of languages which could already benefit from concentrated efforts to create and deploy voice technology. We believe that there is an urgent need to continue with both the development of core speech technologies (such as those described in this report), and with the creation of social-development applications that utilize these technologies. In Africa, where more than 2 000 languages are spoken, this may seem like a daunting task. However, most of these languages fall into closely related families. If the community were to focus on the large regional languages such as Wolof, Swahili and Amharic initially, and use developments in those languages as a platform to create tools and resources for several other languages (with the ordering based on both the practical importance of the language and its relations to these “seed” languages), the impact of voice technologies in Africa can be multiplied in a relatively short period of time.

REFERENCES

[Deliverable D3.1] P Bagshaw, E Barnard and O Rosec, "Report on state of the art and development methodology", Deliverable 3.1 of VOICES project, Sept 2011

[Deliverable D3.2] E Barnard and D van Niekerk, "Language packs for local languages", Deliverable 3.2 of VOICES project, Feb 2013

[Deliverable D3.3] E Barnard, "Report on user acceptance studies – cycle 1", Deliverable 3.3 of VOICES project, May 2012

[Deliverable D3.4] E Barnard, "Report on user acceptance studies – cycle 2", Deliverable 3.4 of VOICES project, May 2013

[VoiceXML] http://www.w3.org/TR/voicexml21, Accessed on 12 June 2013

[CCXML] http://www.w3.org/TR/ccxml, Accessed on 12 June 2013

[PLS] http://www.w3.org/TR/pronunciation-lexicon, Accessed on 12 June 2013

[SSML] http://www.w3.org/TR/speech-synthesis, Accessed on 12 June 2013

[SRGS] http://www.w3.org/TR/speech-grammar, Accessed on 12 June 2013

[SISR] http://www.w3.org/TR/semantic-interpretation, Accessed on 12 June 2013

[Woefzela] https://code.google.com/p/woefzela, Accessed on 12 June 2013

[HTK] http://htk.eng.cam.ac.uk, Accessed on 12 June 2013

[Kaldi] http://kaldi.sourceforge.net, Accessed on 12 June 2013

[Festival Speech Synthesis System] http://www.cstr.ed.ac.uk/projects/festival, Accessed on 12 June 2013