VOIce-based Community-cEntric mobile Services for social development

Grant Agreement Number 269954

Deliverable No D3.2
Language packs for local languages

January 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:	D3.2
Deliverable title	Language packs for local languages
Contractual Date of Delivery:	Sept 2011
Actual Date of Delivery:	4 Feb 2012
Editor (s):	
Author (s):	Etienne Barnard, Daniel van Niekerk
Reviewer (s):	Filipe Cabral Pinto, Anna Bon
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):	19
Keywords:	Language packs, text-to-speech synthesis, automatic speech recognition, Bambara, Bomu

CHANGE LOG

Reason for change	Issue	Revision	Date
Document creation	0.1	Etienne Barnard	21-1-2013
Revision	0.2	Filipe Cabral Pinto	24-1-2013
Release version, including feedback on audio-prompted data collection	1.0	Etienne Barnard	4-2-2013

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

The report describes the development of the language packs that were developed to support the *Radio Marche* and *Tabale* use cases of the VOICES project. In particular, slot-and-filler Text-to-Speech (TTS) systems were developed for Bambara and Bomu, two under-resourced languages spoken in Mali, and a small-vocabulary Automatic Speech Recognition system was developed for Bambara. (In addition, a similar TTS system was developed for Malian French in order to support the use case, but that is not considered part of the current deliverable.)

Various open-source tools were developed to support the creation of these language packs; these include a toolkit that can be used to create slot-and-filler TTS systems efficiently and a platform that can be used for crowd-sourced language data collection in under-resourced environments. The latter platform, called VOCS, has both Web and telephone interfaces and is suitable for the collection, verification and annotation of both text and speech resources. In addition to these software elements, the language packs themselves are also being released as open content, and are packaged in a way that makes them highly reusable.

The performance of the TTS systems was verified with both formal and informal listening tests, whereas the Automatic Speech Recognition system was evaluated in terms of phonetic recognition and two small-vocabulary word recognition tasks. These evaluations confirmed that all components function well, and they have therefore been integrated within the use-case applications, as intended.

Various practical lessons were learnt during the process of language-pack development – most importantly, the need for reusable language resources in various language families that can serve as seeds for the development of language technologies in related languages. It is hoped that the current deliverables will serve in precisely that role in future language-technology development in West Africa and elsewhere.

INTRODUCTION

After extensive research into various use cases of the VOICES project, it was determined that two forms of speech technology would be required:

- Slot-and-filler text-to-speech (S&F TTS) systems, in order to convert text-based communiqués on agricultural prices into a spoken format for telephone access and radio transmission.
- A small-vocabulary automatic speech recognition (ASR) system, to enable automatic processing of user responses to a telephone-based system for arranging meetings in rural areas.

To meet the practical goals of the use cases, the S&F TTS systems had to operate in the languages Bambara (also known as Bamanankan) and Bomu (which are widely spoken in Mali), as well as Malian French. The ASR system was required to recognize speech in Bambara. Thus, four language packs were required: TTS in three languages and ASR in one language. The current report summarizes the processes that were followed in order to develop those language packs.

Below, we first describe the steps that were taken to create the basic generic resources that are required for speech technology in any language: grapheme sets, phoneme sets and pronunciation lexicons. The rest of the report then discusses TTS and ASR system development in turn, and in the concluding section we highlight some of the lessons that have been learnt.

Generic Resources

Before the development of speech technology in a target language can commence, a set of basic resources must be in place. These may include some or all of the following:

- The set of letters or symbols used in the writing system of the language – together, these are known as the "grapheme set".
- The set of basic sounds that constitute the spoken form of the language, also known as the "phoneme set".
- A set of common words or phrases.
- A mapping between the written and spoken forms of the language, which is usually given in the form of phonemic transcriptions of common words or pronunciation rules.

For well-resourced languages, these elements are typically widely available and easily accessed. Bambara, with its approximately 20 million speakers, is also relatively well represented in the literature, and we were able to extract phoneme sets, grapheme sets and pronunciation rules from [Morales, 1996] as well as online resources, especially http://en.wikipedia.org/wiki/Bambara_language. Substantial text resources do not seem to be widely available, and we therefore asked informants in Mali to point us to Web sites that contain substantial amounts of Bambara text. These sites (such as the blog http://fasokan.com) were crawled for text, and French and English portions were removed with automatic scripts. The remaining sentences were then corrected manually by a first-language Bambara speaker, and used as our main text resource for technology development.

Since Bomu is a much smaller language (with a few hundred thousand speakers, according to http://www.ethnologue.com/show_language.asp?code=bmq), comparable resources are much harder to come by. Fortunately, we only intended to create an S&F TTS system in Bomu, and our linguistic informants in Mali indicated that the mapping between writing and speech in Bomu is straightforward. We therefore used a first-language Bomu speaker to translate the potential contents of the communiqués (carrier phrases, content words and relevant number ranges) from French into Bomu – as we describe below, these resources were sufficient for the creation of the Bomu TTS language pack.

TEXT TO SPEECH SYSTEMS

The slot-and-filler TTS systems in Bambara and Bomu were implemented using a limited domain unit-selection approach using word-sized units by customising existing open-source TTS software. These systems were then augmented with a simple REST interface and deployed on the Emerginov platform for integration into the greater Radio Marché system. Evaluations of the resulting synthesised messages were done and further development was done to improve synthesis quality and speed based on feedback from this process.

The following is a summary of activities:

1. Customising existing TTS software to implement unit-selection based on word-sized units for synthesis audio from text messages necessary for this task.
2. Implementing basic text-processing modules for Bambara and Bomu to automatically convert audio recordings and orthographic transcriptions to acoustic inventories during development and handle text input at synthesis time.
3. Developing and evaluating a prototype system followed by development of final systems.
4. Deploying these systems on Emerginov for integration into the larger system, piloting and documenting the software and systems developed.

These tasks are described in more detail in the following sections.

System design and implementation

The basic requirement for the TTS component of the system was the generation of a limited set of pre-defined messages with systematically varying content including dates, quantities, and people and place names. This had to be developed given a limited set of recordings minimally covering the required base messages, lists of names and number sets. For this purpose, it was decided to build limited domain unit-selection TTS systems based on word-sized acoustic unit inventories. This approach would re-use existing software for managing audio data, synthesising messages and processing text input (Figure 1). In addition, tools for automatically segmenting speech and constructing the acoustic inventory from audio and transcriptions could be used directly to minimise manual effort and thus construction time and costs (Figure 2).

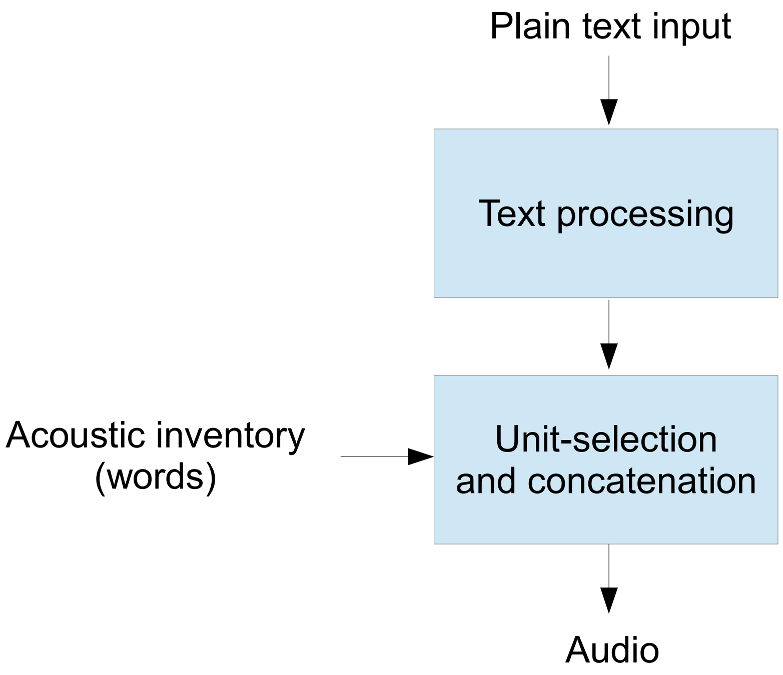

Figure 1: TTS Synthesis process

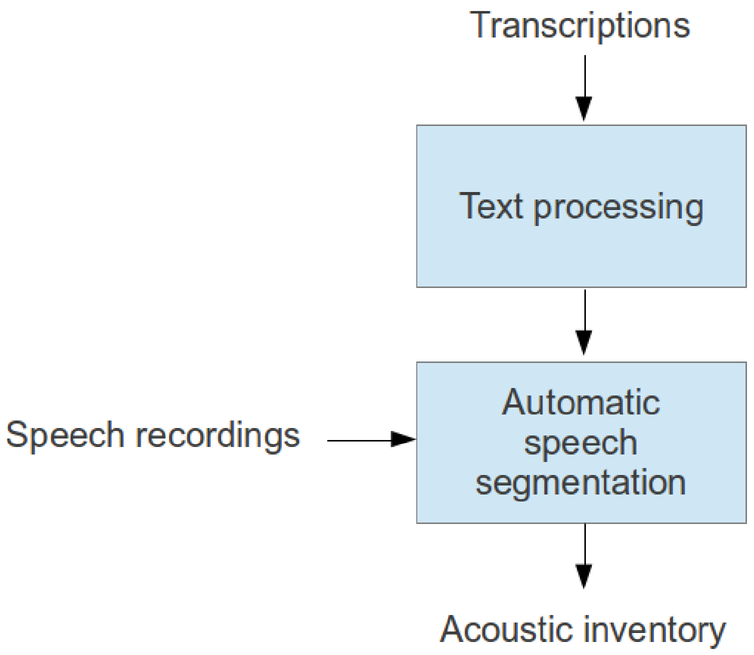

Figure 2: TTS Development process

Given this approach, existing software was customised to support word-sized units and the main task involved developing suitable text-processing modules for Bambara and Bomu as described below.

The following text-processing modules and resources were developed for Bambara and Bomu:

- Basic phoneme sets and letter to sound rules in order to use automatic speech segmentation software as illustrated in Figure 2.
- Text normalisation routines to perform number and date expansions.

Speech recordings were received from Mali in unprocessed form, requiring the following preparation before the process in Figure 2 could be applied:

1. Utterance chunking.
2. Manual transcription corrections and verification.
3. Audio processing including noise reduction and removal of reverberation.

The Bambara system was completed first as a prototype and evaluated informally to determine suitability. The speech output quality was judged to be acceptable with minor feedback asking for longer pauses between messages and dynamic content. The Bomu system was developed subsequently with recordings done at a slower speech rate. The systems were also adapted so that pauses could be inserted between words.

The final systems were then deployed on Emerginov and integrated into Radio Marché as described below.

System deployment and integration

For integration into the Radio Marché system, the TTS systems had to be deployed on the Emerginov web platform. This platform consists of a Linux-based system running Apache and PHP. The runtime systems were largely implemented in Python using Numpy and Scipy with one module written in Perl and were required to present a REST interface to which a text query could be sent, returning an audio file in RIFF Wave format. This interface was written in PHP. As the Radio Marché platform generated messages in a JSON format, a conversion script was also required to convert this into plain text for querying the TTS systems. This process is summarised in Figure 3.

After system deployment and testing two further improvements were made to increase the system usability in Mali:

- Optimisation of audio size using 8-bit u-law encoding of audio.
- Increasing of synthesis speed by compiling parts of the system using Cython.

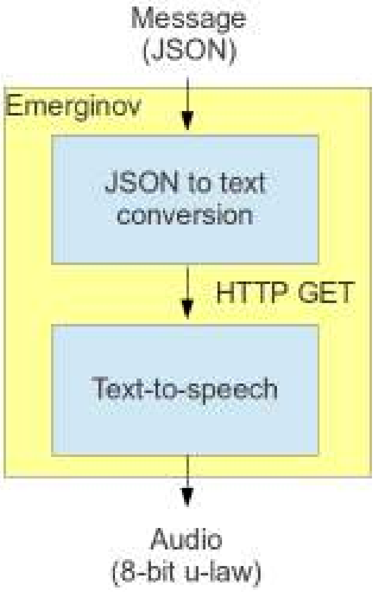

Figure 3: TTS system deployment

AUTOMATIC SPEECH RECOGNITION SYSTEM

As was summarized in Deliverable 3.1, modern approaches to Automatic Speech Recognition (ASR) employ statistical techniques, using Hidden Markov Models (HMMs). As is shown in Figure 4 below (and discussed in more detail in Deliverable 3.1) such ASR systems contain several language-independent components and three principal language-specific modules: the acoustic model, pronunciation dictionary and language model. Our goal was to develop a basic ASR system in the Bambara language to support the *Tabale* event organiser – hence, we were required to develop each of these modules for Bambara.

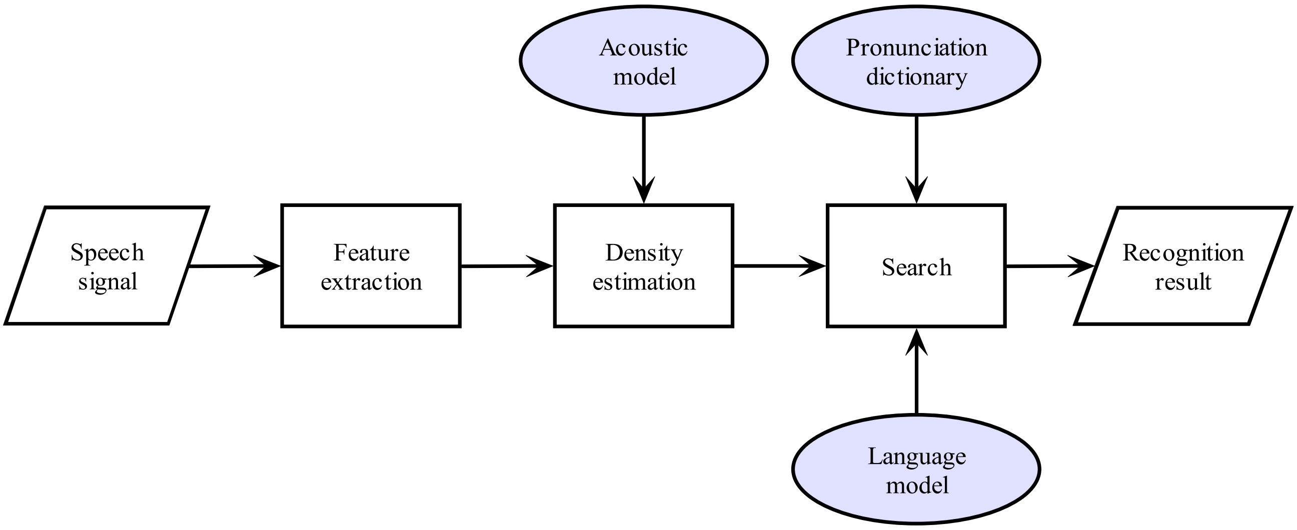

Figure 4: Block diagram of a basic speech-recognition system. Language-specific components are shaded.

Bambara is a language with a highly regular orthography (that is, the writing system creates a close correspondence between the spelling and pronunciation of standard words in the language). It has become increasingly clear in recent years that grapheme-based systems are highly competitive in such languages [Kanthak and Ney, 2003]. In grapheme-based systems, acoustic models are created for each of the *letters* (rather than *phonemes*) in a language, thus removing the need for an explicit pronunciation dictionary. Basson and Davel [2012] have shown that grapheme-based systems may have problems with certain categories of words (such as foreign words or abbreviations), but can outperform phoneme-based systems on the standard words in a language. Since the *Tabale* application will require recognition of standard words only, we have implemented a grapheme-based ASR system, and did not require a pronunciation dictionary.

In Deliverable 3.1 we discussed the distinction between rule-based and statistical language models, and pointed out that rule-based approaches are preferable for small, constrained tasks. Recognition in *Tabale* is exactly such a task; hence, a rule-based language model was implemented. Using a text editor, the expected utterances in the task were specified in the

Backus-Naur form for regular languages; the resulting text file is processed with standard ASR tools for use in the run-time ASR system.

Development of the third language-specific component, namely the acoustic model, was considerably more complicated. Since modern approaches to acoustic modelling employ statistical methods, we required a representative sample of speech in Bambara. To be used in a current ASR system that speech must also be transcribed orthographically (that is, the words corresponding to each of the recordings must be entered into a text file). For acceptable speaker independence, speech must be recorded from at least 20 different speakers [Barnard, 2009], with an approximately equal split between male and female speakers. Also, the dialects of the speakers should match the dialects that will be encountered in the application of the ASR system as closely as possible; since *Tabale* is being deployed in Mali, we therefore required that the speech be collected from Bambara speakers in that country.

Audio-prompted data collection in Bambara

As discussed in [De Vries, 2011], there are two basic approaches to producing such a transcribed corpus:

1. Speakers can be allowed to speak utterances at will, and human transcribers subsequently listen to the recordings and produce the corresponding orthographic transcriptions.
2. Speakers can be prompted to produce specified utterances; human verifiers or automated means can then be used to verify (and possibly correct) the prompts as transcriptions of the recordings.

The second method requires substantially less human effort, and has been found to be quite effective for the development of ASR systems even when only automated verification of utterances is employed [Davel, 2012]; we have therefore decided to use such a prompted approach for the Bambara ASR system.

Prompting is usually done in printed format. However, the difficulties of distributing printed prompt sheets in Mali, along with the possibility that participants could be partially or completely illiterate, required us to develop an alternative approach. Thus, we have decided to experiment with audio prompting, in which the respondent is asked to repeat a prompt that is spoken over the telephone.

Collecting a corpus with the limited infrastructure available in Mali required an innovative approach, and we therefore developed a data-collection platform called VOCS which is suitable for crowd-sourced data collection in a resource-constrained environment. The full functionality of VOCS allows for various forms of data collection. VOCS has two interfaces: a Web interface (which can be used by authenticated workers and administrators) and a telephone interface (for use by the workers). Administrators are responsible for managing the various phases of crowd-sourced corpus creation: worker registration, the management of text sentences and audio recordings, uploading of prompting material, etc. Workers, on the other hand, can create sentences and recordings, validate existing content, or provide annotations.

For our purposes, the most important capability is that VOCS supports the collection of audio-prompted speech data in a crowd-sourcing style. That is, the call-flow in Figure 5 (translated into the appropriate language – Bambara in our case) is used to solicit prompted utterances from callers; these utterances are then saved in a database along with relevant metadata (such as the prompt that had been played, the time of the call and the telephone number from which the call had been received).

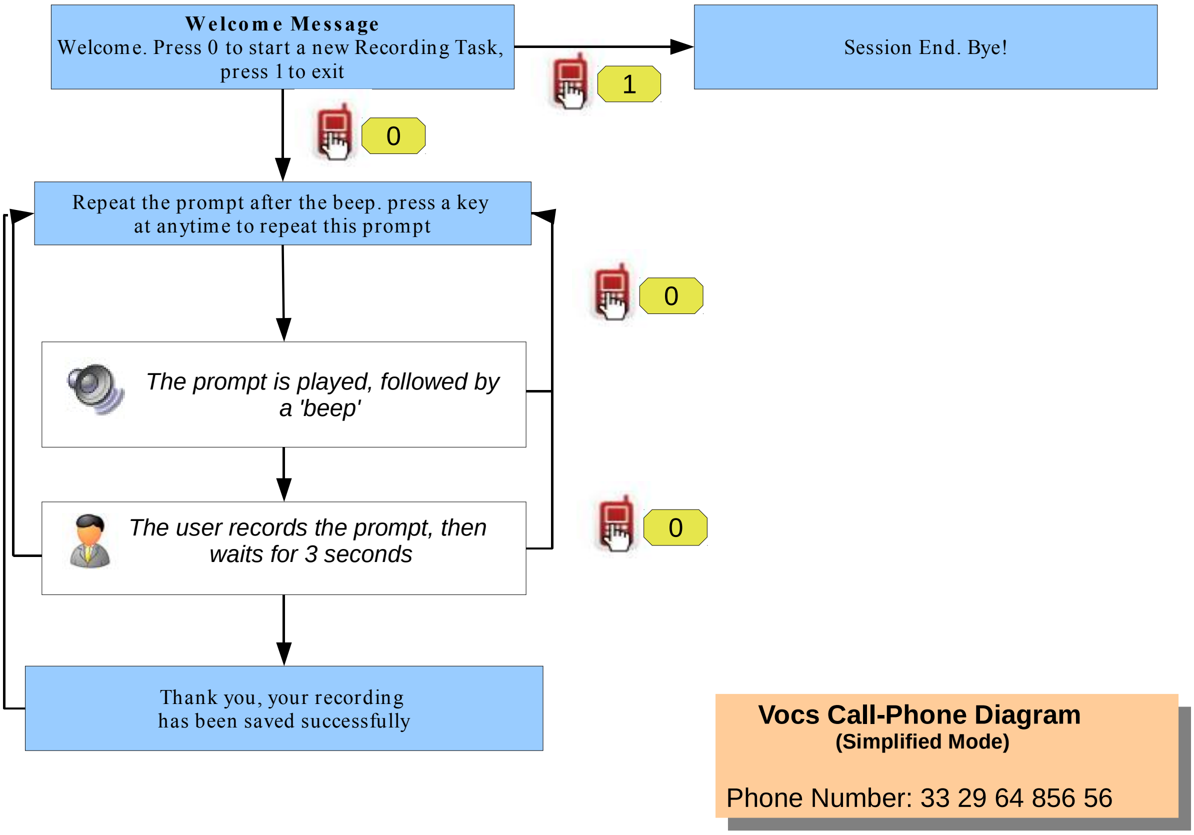
Welcome Message
Welcome. Press 0 to start a new Recording Task, press 1 to exit
Session End. Bye!
1
0
Repeat the prompt after the beep. press a key at anytime to repeat this prompt
0
The prompt is played, followed by a 'beep'
The user records the prompt, then waits for 3 seconds
0
Thank you, your recording has been saved successfully
Vocs Call-Phone Diagram
(Simplified Mode)
Phone Number: 33 29 64 856 56

Figure 5: Call flow of voice-prompted data-collection application.

Using this approach, and the assistance of paid field workers in Mali, we were able to collect approximately 2 500 utterances from approximately 50 Bambara speakers. This corresponds to a corpus of about 2.5 hours of transcribed speech, which is smaller than we had hoped to collect. Fortunately, our previous experience with such corpora suggests that useable small-vocabulary ASR can be achieved with appropriate system design [Van Heerden, 2009].

A number of practically important lessons were learnt during the data-collection process. In particular, even though the majority of respondents found the IVR application easy to use, the amount of speech that could be collected during a typical 30-minute session from a single caller was limited by a number of factors:

- Callers found the audio prompts too long to remember comfortably. Even though this factor had been considered when the corpus of sentences was being designed (and the shortest sentences from the available text corpora had been selected as prompting material), callers often had to listen to prompts repeatedly in order to repeat the sentences accurately.

- Some sounds in the corpus were difficult for speakers of the local dialect to pronounce accurately – callers reported that the voiceless bilabial plosive /p/ and the voiceless labio-dental fricative /f/ were problematic in this regard.
- The voice in which the prompts had been recorded (for repetition by the callers) was sometimes too rapid and unclear.

As a consequence of these factors, it took an average of approximately 40 seconds per recorded prompt in the field, whereas our in-house experiments and earlier experience with text-prompted corpora had lead us to expect an average of 8-10 seconds per prompt. Based on this experience, a number of suggestions were made by our field workers and respondents for more efficient collection of similar corpora in future:

- Shorter sentences (or sentence fragments) should be used for prompting.
- Greater attention should be paid to the creation of the initial prompts; for example, the voice artist recording those prompts should study them in detail before proceeding to the recording phase, in order to ensure that the semantics of the sentences are clear before they are rendered.
- In the field, it would be useful if respondents could familiarize themselves with the prompts to be recorded before the session starts – for literate respondents, text-based versions could be made available, or if that is not feasible, sentences could be played to respondents off-line before initiating the recording session.

By incorporating these suggestions, we should be able to create somewhat larger speech corpora with similar effort in future work.

Training acoustic models

As mentioned above, the vast majority of modern ASR systems employ Hidden Markov Models (HMMs) for acoustic modeling, and several open-source tools are available to support their development. We use HTK 3.4 from Cambridge University to build a context-dependent cross-word HMM-based grapheme-based recogniser with triphone models. Each model had 3 emitting states with 8 mixtures per state. (This combination has proved to be robust and accurate in various previous developments [Van Heerden, 2009].) 39 features are used: 13 MFCCs together with their first and second order derivatives. Cepstral Mean Normalisation (CMN) as well as Cepstral Variance Normalisation (CVN) are used to perform speaker-independent normalisation. A diagonal covariance matrix is used; to partially compensate for the implicit assumption of feature independence, semi-tied transforms are applied.

In order to calibrate the performance of these acoustic models for small-vocabulary recognition, various analyses were performed.

- For **grapheme recognition**, each utterance is treated as a sequence of graphemes, and the recognition grammar is simply a "grapheme loop" – that is, any grapheme can be followed by any other grapheme. This is a hard recognition task: on large, broad-band corpora of read speech, accuracies of around 70% are considered state-of-the-art, whereas accuracies of 30% to 35% are typically achieved with large corpora of natural telephone speech. We measured an accuracy of 21.2% on our test set. As expected, this is much lower than the accuracies achieved with large corpora, but substantially higher than chance performance (which would be around 3% with our grammar).

- In **designed-vocabulary isolated word recognition,** we selected a group of 10 words, each of which contained at least 3 graphemes, and created a grammar specifying that each of these words could be spoken in isolation. This represents a typical small-vocabulary recognition task, where the application designer is able to choose keywords that are somewhat distinctive. Our test set now corresponded to all the occurrences of these words in utterances of our test speakers, and we achieved 95.6% recognition accuracy on these test utterances.
- Finally, for **mixed-vocabulary isolated word recognition,** our grammar contained 5 "designed" words (i.e. with length of 3 or more graphemes) and 5 short words (each consisting of only 2 graphemes). This models recognition tasks in which certain words are unavoidably quite short and confusable, given the logic of the application. In this case, we measured a recognition accuracy of 89.3% on the utterances of our test speakers.

Taken together, these results suggest that a useable level of accuracy has been achieved. As would be expected from the limitations of our corpus, the recognizer is considerably less capable than typical systems in well-resourced languages, but with proper application design, ASR accuracies of 90 - 95% for small-vocabulary isolated-word tasks are certainly in the useable range. In addition, the tools that we have created make it possible to collect and incorporate application-specific utterances rapidly and efficiently, so that higher accuracies and larger vocabularies can be achieved if required.

CONCLUSION

The development of the local-language language packs entailed the creation of several open-source or open-content components. These include the following:

1. VOCS, a platform for crowd-sourced content creation in under-resourced languages.
2. A toolkit for the creation of slot-and-filler text-to-speech systems.
3. A wrapper around the open-source HTK toolkit, which makes it possible to develop ASR systems quickly and conveniently.
4. A Bomu S&F TTS system for reading back market prices.
5. A Bambara S&F TTS system for reading back market prices.
6. A Bambara ASR system, suitable for small-vocabulary isolated word or phrase recognition.
7. Several additional Bambara resources, including word lists, manually curated text, and multi-speaker recordings of around 2,500 utterances.

In addition to these tangible deliverables, we have also learnt a number of valuable lessons. Most importantly, we have seen that audio-prompted data collection is a feasible approach to the creation of speech corpora in under-resourced languages. As far as we know, this approach has not been used previously, and it simplifies the logistics of corpus creation considerably. We believe that our current collection protocol can be improved, resulting in more efficient data collection. In particular, we have learnt about the challenges that less-literate callers experienced with our system (as summarized above), and should be able to perform similar collections more efficiently in future.

We have also seen that S&F TTS systems are a straightforward and highly usable solution to the problem of audio-content creation. Although our toolkit greatly simplifies the development process for such systems, it is probably still too optimistic to believe that such systems can be developed without assistance from speech-technology experts.

Finally, our development has again emphasized the difficulties involved in obtaining basic linguistic resources in under-resourced languages. Since languages occur in families, and there are often substantial similarities between related languages, there is a clear need for repositories that collect available resources and organize these resources in ways that make it easy to create novel resources from those developed for related languages.

REFERENCES

Barnard E, Davel M, and Van Heerden C, "ASR corpus design for resource-scarce languages," in Proc. Interspeech, pp. 2847–2850, 2009.

Basson WD and Davel MH, "Comparing grapheme-based and phoneme-based speech recognition for Afrikaans," in Proc. PRASA, pp. 144-148, 2012.

Davel M, Van Heerden C and Barnard E, "Validating Smartphone-Collected
Speech Corpora" in Proc. SLTU, pp 68–75, 2012.

De Vries N, Badenhorst J, Davel M, Barnard E and De Waal, A. "Woefzela — An Open-Source Platform for ASR Data Collection in the Developing World" ," in Proc Interspeech , pp. 3177-3180, 2011

Kanthak S and Ney H, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in Proc. ICASSP, pp. 845–848, 2002.

Morales J, "J'apprends le Bambara", Editions ACCT - KARTHALA, France, 1996

Van Heerden C, Barnard E, and Davel M, "Basic speech recognition for spoken dialogues," in Proc. Interspeech, pp. 3003–3006, 2009.