Incorporating Speech Synthesis in the Development of a Mobile Platform for E-learning

J.C. Roux[1], P.E. Scholtz[2], D. Klop[2], C. Povlsen[3], B. Jongejan[3], A. Magnusdottir[4]

(1) School of Languages, North-West University,
Potchefstroom 2520, South Africa
(2) Stellenbosch University
Victoria Street, Stellenbosch 7600, South Africa
(3) University of Copenhagen
Njalsgade 140-142, 2300 Copenhagen, Denmark
(4) Independent Researcher
justus.roux@nwu.ac.za, pscholtz@sun.ac.za, dk@sun.ac.za, cpovlsen@hum.ku.dk, bart@cst.dk, astaolga@gmail.com,

Abstract

This presentation and accompanying demonstration focuses on the development of a mobile platform for e-learning purposes with enhanced text-to-speech capabilities. It reports on an international consortium project entitled *Mobile E-learning for Africa* (MELFA), which includes a reading and literacy training component, particularly focusing on an African language, isiXhosa. The high penetration rate of mobile phones within the African continent has created new opportunities for delivering various kinds of information, including e-learning material to communities that have not had appropriate infrastructures. Aspects of the mobile platform development are described paying attention to basic functionalities of the user interface, as well as to the underlying web technologies involved. Some of the main features of the literacy training module are described, such as grapheme-sound, correspondence, syllabification-sound relationships, varying tempo of presentation. A particular point is made for using HMM (HTS) synthesis in this case, as it seems to be very appropriate for less resourced languages.

1. Introduction

Mobile technology development opens new doors for education, particularly in developing countries. E-learning possibilities have over the last few years been substantially enhanced by the development of mobile applications that can either support existing learning methods, or become an autonomous and active tool in the learning process. The use of mobile phones in, for instance, language learning is proving to be one of the growth areas in this domain: computer-assisted language learning (CALL) is gradually being complemented by mobile-assisted language learning (MALL) (cf. Chinery, 2006). Among the many applications, the launch of BBC`s Learning English program on Nokia's mobile English Language Teaching (ELT) platform in China in 2007 is a case in point, (cf. indiantelevision, March 2009). While it is true that internet penetration in the African continent still is extremely limited, reaching only 5.6% of its population in relation to a penetration rate of 26.5% for the rest of the world in 2008 (cf. internetworldstats, April 2009), it is also true that Africa has become the fastest growing mobile market in the world. In 2007 there were 280.7 million mobile subscribers in Africa, representing a penetration rate of 30.4% (cf. whiteafrican, March, 2009). Given this situation, it has become clear that smartphones (being an indirect entrance to the Internet) are fast becoming the PCs of the developing world. This is exemplified by a wide range of innovative applications that are being developed for mobile platforms e.g. the use of handheld computers to enhance the teaching practices of teachers in disadvantaged rural communities in South Africa (Power & Sankale, 2007). 32% of South African adults are functionally illiterate (Aitchison & Harley, 2006), in other words their reading and writing skills are insufficient for their ordinary practical needs. Given this high illiteracy rate in Africa, but simultaneously also a wide penetration rate of mobile phones, it stands to reason that implementing this technology and enhancements thereof will breach communication borders and address literacy challenges. Literacy, being the first step towards reading and learning, is an area in need of development at higher technological levels. A prominent aspect of literacy development and language learning is having direct access to the sounds of the language involved. Hence, in principle, any learning system that provides access to sound (either as input or as output) could significantly enhance the learning process and circumvent the reading and writing problems experienced by functionally illiterate adults (Lumsden, Leung & Fritz, 2005). Mobile technology can offer support to persons with limited literacy skills as well as expand and enhance their existing reading and writing skills through opportunities for experiential learning through touch-screen based interaction.

This paper describes a subsection of an ongoing international consortium project entitled *Mobile E-Learning for Africa* (MELFA), which integrates a text-to-speech (TTS) engine as part of the e-learning platform, (cf. www.melfaproject.net). The MELFA project focuses, *inter alia*, on communicative aspects related to South African English as well as to an indigenous African language, isiXhosa, which is one of the eleven official languages of the country. Although concatenative speech synthesis normally provides high quality speech output (which is needed for applications such as these), the process is relatively complicated and time consuming. Given the challenges of eventually developing systems for all the local languages, it was necessary to look into alternative approaches that could also render high quality speech, and possibly cut down on the development time.

In particular, this presentation describes the implementation of an HMM-based text-to-speech synthesis system (HTS) in a module devoted to the development of literacy and reading skills in isiXhosa. This module is developed on top of a generic web-based, speech-enabled platform.

2. Platform Development

Although the MELFA project is primarily intended to deliver multilingual, mobile e-learning applications with smartphones as target, the same concept can be applied to desktops, notebooks, netbooks and feature phones of the future. This is a pilot project assessing, *inter alia*, the efficacy of implementing a text-to-speech system in a mobile literacy development programme for an African language. Although the results of usability tests are not yet available, and we can't report on that, we nevertheless provide an overview of the development of the mobile platform as well as the application of the text-to-speech system in a literacy training programme. Whilst initially focusing on English and isiXhosa, the aim is to develop a generic platform that supports the rapid development of any application that can benefit from synthetic speech output.

2.1 Platform Design

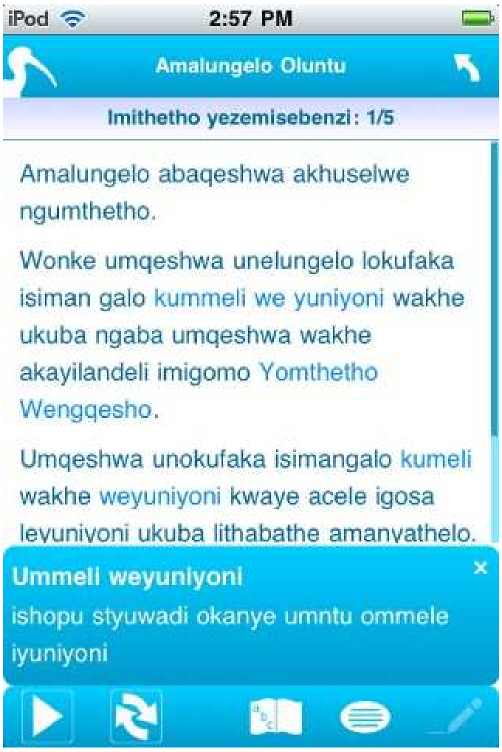
A fundamental requirement of the platform is the on-demand availability of the text-to-speech functionality. Therefore, the speech engine must run natively on the mobile device. In order to support flexibility of textual content and facilitate user generated content, a general domain TTS system is required, capable of effortlessly rendering intelligible speech from unpredictable texts. This issue is discussed in greater detail in the section on speech technology.

The platform presents the user with textual information. All text displayed on the screen can be rendered as speech. With the TTS running natively on the device, visual cues can be used to reinforce the connection between the visual and auditory aspects of the language. An example of this is the synchronised highlighting of words as they are spoken. The same concept of synchronised highlighting can be applied to the various levels of the constructed linguistic hierarchy, i.e. a heterogeneous relation graph, supporting the synthesizer (Taylor, Black & Caley, 2001), e.g. to sentences, phrases, tokens, words and even syllables and phonemes. As this information is embedded within the TTS engine, its visual exposure to the user, in time with the speech, allows for the development of innovative language learning and literacy training applications.

2.2 User Interface Design

Usability is a primary concern to the development of e-learning applications. Traditionally a stylus has been used to interact with Windows Mobile devices. While a stylus is ideal for handwriting and letter recognition, it is not necessarily suited to navigating user interfaces. With the introduction of the iPhone there has been renewed interest around finger friendly touch-based user interfaces. Many device vendors, especially HTC, has followed suit by providing finger friendly shells on top of the antiquated Windows Mobile interface.

In order to provide a very simple and intuitive user interface, custom touch gestures have been implemented. The user can scroll pages using a simple touch and drag gesture, replacing the default Windows Mobile scrollbars. A non-interactive scroll position indicator is displayed, providing information as to the length of the page, as well as the relative position of the current view. The TTS functionality is invoked with a similar gesture. To have a portion of text read aloud the user simply drags a finger from left to right over the text to create a selection. Once a selection is created it can be expanded by moving the finger up or down the screen. As the finger is lifted from the screen, the selection is read aloud. As multilingual content can be displayed in a single page view, each readable segment is tagged with a language code so that the TTS engine can determine which voice to use for which segment. The user can also have the entire page contents read out load by clicking on a play button on the bottom toolbar. A translate button is also provided to cycle between the available languages. See Figure 1.



Amalungelo Oluntu

Imithetho yezemisebenzi: 1/5

Amalungelo abaqeshwa akhuselwe ngumthetho.

Wonke umqeshwa unelungelo lokufaka isiman galo kummeli we yuniyoni wakhe ukuba ngaba umqeshwa wakhe akayilandeli imigomo Yomthetho Wengqesho.

Umqeshwa unokufaka isimangalo kumeli wakhe weyuniyoni kwaye acele igosa leyuniyoni ukuba lithabathe amanyathelo.

Ummeli weyuniyoni

ishopu styuwadi okanye umntu ommele iyuniyoni

Figure 1: A page displaying an isiXhosa definition of a term found in the content.

2.3 Web Technology

The recent proliferation of full HTML browsers on smartphones, coupled with mobile development platforms focusing mainly, or exclusively, on the use of web technologies for application development (cf. developer.apple, March, 2010 and developer.palm, March, 2010), are clear indicators of the web becoming the *de facto* platform on which to deliver innovative applications, not only on desktops, but also on mobile devices. The use of open web technologies, like hypertext markup language (HTML), cascading style sheets (CSS) and JavaScript, is ideal for rapid content and application development, prevents vendor lock-in and maximises deployment opportunities. At the time of writing our UI framework has been ported from Windows Mobile to the iPhone and Android platforms, with little to no effort.

Within this web-based framework administrators can revise, edit and translate content remotely. The updated content can then be synchronized with connected clients, without user intervention. The same concept of real-time updates applies to the application logic itself. The embracing of web technology does not imply that complete cross-browser support is sought, nor does it imply that network connectivity is required. Once the application has been deployed to the device, all content and functionality is available offline.

The client application is built on the principle of the site specific browser, an embedded browser engine without the familiar browser user interface components. Access to speech is provided by plugging the TTS engine directly into the browser engine, exposing its API to the browser's JavaScript environment. On standard browsers, without such a TTS extension, the application can degrade to either cached audio files, or server-side TTS. The former would require an appropriate plug-in to play the audio files; the latter would require a network connection and a plug-in. Both of these fallback strategies have been implemented, but their dependence on 3rd party plug-ins, like Adobe Flash, for audio playback is still problematic on most smartphone platforms.

The open-source WebKit browser engine (cf. webkit.org, March, 2010) has become almost ubiquitous on smartphones due to its portability, standards-compliance and high performance rendering and JavaScript engines in resource constraint environments. For these reasons it is the primary target of our UI framework and applications. However, as our UI framework is built on the Dojo Toolkit, which provides excellent cross-browser APIs, we can readily move our applications to other browsers, including Firefox.

2.4 Speech Technology

As an e-learning development platform in line with the principles of Web 2.0 user generated content, a TTS system that is capable of naturally and intelligibly rendering any text in the target language is critical. The TTS must also be flexible so that speaking rate can be adjusted without affecting output quality.

For the proof of concept prototype two synthetic voices, a South African accented English female voice and an isiXhosa male voice have been developed using HTS technology (Zen et al., 2007). HTS was chosen due to a number of advantages over competing synthesis techniques:

- language independent architecture (Black, Zen, Tokuda, 2007),
- rapid prototyping of new voices (Roux & Visagie, 2007),
- small footprint (1-3MB per voice),
- flexible synthesis parameters,
- it is fast and portable.

The regular phonemic structure of the Bantu languages, including isiXhosa, provides high accuracy in predictability of phoneme sequences from orthographic representations using handwritten letter-to-sound (LTS) rules. Furthermore, many of these rules are shared across language families, e.g. a prototype isiZulu voice has been built using the isiXhosa LTS rule set.

The voices are built using HTS from these handwritten LTS rules and other basic linguistic features, including phoneme, syllable and word position counts (Roux & Visagie, 2007). The high quality of the synthetic speech, especially in terms of naturalness and intelligibility, has been confirmed, informally, by a number of mother-tongue speakers. The male isiXhosa voice was built from only 45 minutes of recorded speech.

For the South African English voice a similar approach was used as can be found in the literature (Tokuda, Zen, & Black, 2002). In this case the text processor is supported by a large pronunciation lexicon and an LTS rule set automatically trained from this lexicon. Besides the addition of syllabic stress, the exact same linguistic feature set was used for this voice as for the isiXhosa voice.

The remarkable adaptability of HTS to novel languages, the relative ease with which new voices can be built and the high quality of these voices makes it an ideal technology for the development of TTS systems in any language, but especially languages with limited linguistic resources.

As the TTS engine is written in ANSI C it is highly portable to all platforms that support native code development, including most smartphone platforms. At the time of writing the TTS engine has been ported to the Windows Mobile and Android platforms.

3. Applications for the Development of Literacy and Reading Skills

Although there are different views on the nature of developing literacy and reading skills, this project follows an adapted interactive reading model (cf. Deschant, 1991) whilst using technology to enhance particular functions.

This approach makes provision for:

- developing phonemic awareness: i.e. linking graphic symbols with sound utterances in a meaningful context;
- developing sight vocabulary by substituting vocabulary in meaningful contexts;
- creating comprehensible short reading passages through personal selection;
- developing writing skills;
- developing open reading support environments.

These functions are described in more detail below.

The phonological approach, i.e. working with phonemes and grapheme-to-phoneme representations, is useful and appropriate at the decoding/word recognition levels of reading; suitable for persons with limited levels of literacy; persons learning to read for the first time, or in a new language with a different phonemic system. Based on this established observation a phonemic awareness functionality has been integrated into the platform. This learning feature is implemented by letting the system display a number of predetermined sentences representing all the phonemes and phoneme combinations in the language; these sentences are generated randomly by the system. The particular sentence is presented in large legible fonts with an option for the user to listen to it as normal, slow or fast renditions. In order to focus the user's attention on the temporal nature of speech, synchronised spoken word highlighting is implemented. This can be taken a step further as the whole utterance can optionally be syllabified, rendering a one to one representation of syllables and sounds, thus reinforcing the sound and (orthographic) symbol relationship. An African language such as isiXhosa presents itself very well to this approach as it displays a very regular consonant-vowel /CV/ syllabic pattern as indicated in the example below:

(1) Graphic input: **Abafana badlala egadini**. ("The boys are playing in the garden.")

- TTS generates 'normal', 'slow' and 'fast' renditions on request.
- Syllabification with optional tempo variations: **A ba fa na ba dla la e ga di ni**

In order to expand reading and writing skills for illiterate users, entries from sight vocabulary have been selected and stored in the system. The learning idea that has been implemented is realised by optionally allowing for a particular word to be substituted in the sentence, e.g.

(2) Graphic input: **Abafana badlala entabeni**. ("The boys are playing at the mountain.")

A list of isiXhosa words can be accessed independently as substitutions with the possibility of generating pronunciations for each item via the TTS system.

Short reading passages (paragraphs) in isiXhosa are provided by the system to the user to develop comprehension skills in the language. Thus, a user will select the whole paragraph or sections thereof by a touch screen gesture and listen to the TTS generated speech. This can be done at different speaking rates and is followed by a set of multiple choice options eliciting the correct (semantic) interpretation. These choices are also presented graphically as well as auditorily to the user.

Literacy training also involves developing **writing skills**. As part of this training function, the system prompts the user auditorily to type a particular word - in the case of a correct transcription the word is repeated by the system. In the case of the user not being able to type the word, selected cues are graphically provided in support. The focus is to familiarise the user with the orthographic and aural aspects of the language.

4. Conclusion

Advances in TTS technologies have made it possible to enhance the functionalities of mobile e-learning content significantly. HMM driven speech synthesis in particular represents a relatively easy approach to rapidly provide applications for less resourced languages; this includes applications in the domain of language and literacy training.

5. References

Aitchison, J.J.W. & Harley, A. 2006. South African illiteracy statistics and the case of the magically growing number of literacy and ABET learners. Journal of Education, No. 39, pp. 89-112.

Black A., Zen H., Tokuda K. 2007. Statistical Parametric Speech Synthesis. Proceedings of ICASSP, 2007, pp.1229-1232.

Chinery, G.M., 2006. "Emerging technologies; Going to the MALL: Mobile Assisted Language Learning", Language Learning & Technology, Volume 10, Number 1, pp. 9-16.

Deschant, E. 1991. Understanding and teaching reading: An interactive model. Hillsdale, NJ; Lawrence Erlbaum.

Lumsden, J., Leung, R. & Fritz, J. 2005. Designing a Mobile Transcriber Application for Adult Literacy Education: A Case Study. In: Proceedings of the International Association for Development of the Information Society (IADIS) International Conference Mobile Learning 2005.

Power, T. & Sankale, J. 2007. In the palm of your hand: supporting rural teacher professional development and practice through the use of mobile phones and other handheld digital devices. In: Meraka Innovate Conference, April 18-20 2007, Pretoria, South Africa.

Roux J.C., Visagie A.S. 2007. Data-driven approach to rapid prototyping Xhosa speech synthesis. Proceedings of ISCA SSW6, Bonn, 143-147.

Taylor P., Black A.W. & Caley R. 2001. Heterogeneous relation graphs as a formalism for representing linguistic information. Speech Communication, Volume 33, Number 1, pp. 153-174.

Tokuda K., Zen H., Black A.W. 2002. An HMM-based speech synthesis system applied to English. Proceedings of the IEEE Workshop on Speech Synthesis, pp. 227-230

Zen H., Nose T., Yamagishi J., Sako S., Masuko T., Black A.W., Tokuda K. 2007. The HMM-based Speech Synthesis System Version 2.0. Proceedings of ISCA SSW6, Bonn, 294-299.

Internet resources:

http://www.indiantelevision.com/headlines/y2k7/oct/oct3 php, accessed on 2009-03-20.

http://www.internetworldstats.com, accessed on 2009-04-07.

http://whiteafrican.com/2008/08/01/2007-african-mobile phone- statistics, accessed on 2009-03-19.

http://developer.apple.com/safari/, accessed on 2010-03-12.

http://developer.palm.com/, accessed on 2010-03-12.

http://webkit.org, accessed on 2010-03-12

http://www.sil.org/lingualinks/Literacy/ReferenceMateri als, accessed on 2008-11-13.