Speech Technology for Information Access: a South African Case Study

Etienne Barnard and **Marelie H. Davel** and **Gerhard B. van Huyssteen**
Human Language Technologies Research Group
Meraka Institute, CSIR
South Africa

Abstract

Telephone-based information access has the potential to deliver a significant positive impact in the developing world. We discuss some of the most important issues that must be addressed in order to realize this potential, including matters related to resource development, automatic speech recognition, text-to-speech systems, and user-interface design. Although our main focus has been on the eleven official languages of South Africa, we believe that many of these same issues will be relevant for the application of speech technology throughout the developing world.

Index Terms: spoken language technology for development, resource-scarce languages

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 prices – or simply do not have access to information that could substantially enhance their quality of life. Spoken dialog systems (SDSs) have the potential to become a primary source of such information in the developing-world context (Tucker and Shalonova 2004). Although traditional computer infrastructure is often scarce in the developing world, telephone networks (especially cellular networks) are spreading rapidly, thus creating tremendous opportunities for speech-based information access. In addition, speech-based systems may empower illiterate or semi-literate people, the vast majority of whom live in the developing world.

Spoken dialog systems can play a useful role in a wide range of application environments, including educational settings, kiosks, embedded devices and more. Hopefully, each of these environments will benefit from the development of speech technologies, but our expectation is that the largest impact will stem from their use in telephone-based information systems. Significant benefits can be envisioned if information is provided in domains such as agriculture (Nasfors 2007), health care (Sherwani et al. 2007; Sharma et al. 2009) and government services (Barnard, Cloete, and Patel 2003).

However, the deployment of such systems faces a number of logistical, technical and financial hurdles. Each of these categories of hurdles warrants serious attention, but in the current contribution we focus almost exclusively on the technical issues. From the perspective of speech technology, 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.
- The development of speaker-independent automatic speech recognition (ASR) systems that function reliably in the local languages of the developing world.
- The development of text-to-speech (TTS) systems that are easily understood in these same languages.
- The design of spoken interfaces that are usable and friendly in diverse, multilingual cultures, by users with limited or no computer literacy.
- The development of tools that support these activities and of platforms that make it possible to deploy them in both experimental and customer-facing applications.
- The selection of application domains that are practically suitable in terms of factors such as economics, the availability of information and marketability.

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.

Below, we briefly summarize and motivate the way we approached each of the above challenges. The majority of this work took place in the context of two large projects aimed at telephone-based information access, namely (a) the OpenPhone project (Sharma et al. 2009), which developed and piloted an information service for caregivers of people living with HIV/Aids in Botswana, and (b) the Lwazi project

(Meraka-Institute 2009), which developed speech technologies in all the official languages of South Africa, and piloted their use in applications aimed at community assistance.

Developing speech technologies for the languages of South Africa

Most of our work has been focused on the languages of South Africa. In this section, we describe our research and development related to these languages. Many of the lessons learned are likely to be widely applicable in the developing world, whereas others are more specific to the South African environment.

Gathering linguistic knowledge

For the languages of the developed world (such as English, German or Japanese), the linguistic knowledge necessary for the development of speech technology is generally well established and easily accessible. The situation in the developing world is much less uniform, ranging from the many languages for which no writing system has been developed, through those such as Swahili or Yoruba which have benefited from extensive linguistic research, to languages such as Hindi and Mandarin, which are at least as well studied as most developed-world languages. The languages of South Africa tend to be around the middle of this spectrum (with the exception of English, which is an official language of South Africa, but is generally excluded from the discussion below because of its world-language status).

South Africa has eleven official languages, of which nine belong to the Southern Bantu (SB) family, the other two being Germanic languages (see Table 1). Many of these languages are also spoken in other Southern African countries, or closely related to such languages. For example, Siswati, Setswana and Sesotho are the largest languages in Swaziland, Botswana and Lesotho, respectively, while Xitsonga is spoken by a large population in Mozambique, and related to other significant languages of that country. As is often the case in the developing world, a local variant of a colonial language (in this case, English) tends to be the primary language for commerce and government, although a large part of the population have little or no mastery of it.

The SB languages are all tone languages. Although a significant body of research has explored aspects of these tone systems (see (Zerbian and Barnard 2008b) for a summary), our understanding of their tonal phonology and phonetics is still not sufficiently concrete to be useful for the purposes of technology development. In our research group, we are therefore involved in several studies aimed at the empirical investigation of these tone systems – initial work was reported in (Zerbian and Barnard 2008a).

Usable descriptions of the phoneme sets of all these languages have been published. As pointed out in (Barnard and Wissing 2008), these descriptions are generally not based on strong empirical foundations, and some revision may be required for the purposes of speech technology. We have nevertheless been able to create a unified phoneme set across the eleven official languages (Meraka-Institute 2009), which will serve as a starting point for ASR and TTS development, while constantly being refined in the foreseeable future.

Language	code	# million speakers	language family
isiZulu	Zul	10.7	SB:Nguni
isiXhosa	Xho	7.9	SB:Nguni
Afrikaans	Afr	6.0	Germanic
Sepedi	Nso	4.2	SB:Sotho-Tswana
Setswana	Tsn	3.7	SB:Sotho-Tswana
Sesotho	Sot	3.6	SB:Sotho-Tswana
SA English	Eng	3.6	Germanic
Xitsonga	Tso	2.0	SB:Tswa-Ronga
Siswati	Ssw	1.2	SB:Nguni
Tshivenda	Ven	1.0	SB:Venda
isiNdebele	Nbl	0.7	SB:Nguni

Table 1: *The official languages of South Africa, their ISO 639-3:2007 language codes, estimated number of home language speakers in South Africa (Lehohla 2003) and language family (SB indicates Southern Bantu).*

Collecting basic resources

Modern speech technology relies heavily on phone(me)-based systems – hence, pronunciation dictionaries (or, equivalently, letter-to-sound rules) are the cornerstone for such technology. In a series of papers, we have developed a bootstrapping approach that supports the accelerated development of such dictionaries, while limiting the level of linguistic expertise required ((Davel and Barnard 2008) and references therein). This approach, which is implemented in open-source software, has enabled us to develop reasonably accurate pronunciation dictionaries for all the South African languages (Davel and Martirosian 2009).

For the development of these dictionaries, we needed lists of frequent words in each of the languages of interest. Such lists were created from textual material obtained by a combination of Web crawling (Botha and Barnard 2005) and contracting with established publishers for use of their copyrighted texts. (The use of additional copyrighted textual resources was unfortunately necessary, given the limited amount of clean, usable data that could be obtained via Web crawling for some of the lesser spoken languages.)

The availability of pronunciation dictionaries makes it possible to extract phonetically balanced selections from the aforementioned text corpora. These phonetically balanced sub-corpora are useful for the specification of speech corpora for the purposes of ASR and TTS development, as we detail below.

Developing ASR

Most modern speech recognition systems use statistical models which are trained on corpora of relevant speech (i.e. appropriate for the recognition task in terms of the language used, the profile of the speakers, speaking style, etc.). This speech generally needs to be curated and transcribed prior to the development of ASR systems, and speech from a large number of speakers is generally required in order to achieve acceptable system performance. In the developing world, where the necessary infrastructure such as computer networks, as well as first language speakers with the relevant training and experience, are limited in availability, the collection and annotation of such speech corpora is a significant hurdle to the development of ASR systems.

Fortunately, the class of SDSs envisioned for information access in the developing world does generally not require the large-vocabulary, natural-language processing capabilities which necessitate such large training corpora (Plauché et al. 2006; Sherwani et al. 2009). For many of these applications, dialogs can be designed that limit the active vocabulary at any point in the interaction to a dozen or fewer words. (Such limited systems are also a useful way to bootstrap systems with larger vocabularies and higher accuracies, since they can be deployed in usable applications that simultaneously perform data collection.) We have therefore investigated the performance that can be achieved with ASR systems that use relatively small corpora (fewer than 200 speakers, less than 10 hours of speech per language). Our research utilizes such a corpus of telephone speech (Barnard, Davel, and van Heerden 2009), based on the phonetically balanced sentences described above, and aimed at the class of applications described in the introduction.

Using the open-source toolkit HTK to train standard context-dependent hidden Markov models (HMMs), we obtain accuracies in the range 54% to 67% for unconstrained phone recognition with a flat language model on the eleven languages of interest (van Heerden, Barnard, and Davel 2009). This is comparable to reported recognition rates with established corpora. These accuracies translate to reasonable to good accuracy on the type of small-vocabulary task (around 10 words) we envision for spoken-dialog systems: error rates ranging between 2% and 12% were measured on such a task (van Heerden, Barnard, and Davel 2009).

A potentially significant observation made during this research is that the number of speakers required to achieve acceptable accuracy is actually much less than the approximately 200 speakers per language who contributed to our corpora. It seems as if 50 speakers per language, and possibly as few as 30 per language, would yield equally good results – if each speaker produces a larger amount of speech (Barnard, Davel, and van Heerden 2009). This tendency was observed for both the Bantu and Germanic languages studied, and could reduce the burden of ASR corpus development significantly if confirmed. We are currently busy with additional experiments to assess the validity of this hypothesis, using speech corpora that are better known internationally.

Developing TTS

State-of-the-art TTS systems rely on phonetically aligned corpora of speech by a single speaker, from which speech segments are excised (for concatenative synthesis) or statistical models estimated (for statistical synthesis). In the developing world, an important practical principle is to minimize the effort and expertise needed to create such corpora. In practice, this implies the use of corpora that are as small as possible, and the use of automatic alignment methods wherever possible. Unfortunately, these two goals are at cross purposes: for smaller corpora, automatic alignments are expected to be less accurate.

Much of our work on TTS therefore focuses on methods that limit manual interaction and maximize alignment accuracy when aligning small corpora in new languages. In a study involving three small TTS corpora (durations ranging between 20 minutes and 46 minutes of speech), we have reached the following conclusions (van Niekerk and Barnard 2009):

- Even for such small corpora, HMM-based alignment is both more accurate and more robust than dynamic time warping (DTW) alignment using phone mapping to an existing TTS voice.
- For these small corpora, the normal "flat start" initialization of HMMs is significantly inferior to approaches that initialize the models to more reasonable values. We have experimented with two approaches: manual alignment of a small subset of sentences (around 20 selected sentences) and broad-category alignment using a well-trained existing recognizer in another language (typically, English). The former approach is somewhat more accurate, but at the cost of requiring some manual alignment; cross-language broad-category alignment is therefore an acceptable option when such alignment is a significant obstacle.
- The optimal HMM parameters for this task are somewhat different than those that have become standard for speech recognition: we find that a single mixture with diagonal covariance matrix works well, but with a larger number of states per model and higher frame rates (5 ms frame rates and 10 ms windows were found to work well).

With these refinements, we are able to perform automatic alignments that are comparably accurate to those created by our non-expert transcribers. Using the techniques described here, concatenative TTS systems were developed for all eleven South African languages. Initial results were very positive: highly intelligible synthesis was achieved in all languages, with minimal manual intervention. Formal perceptual evaluations have also been completed for some of the systems, which confirmed initial results (Meraka-Institute 2009).

Designing spoken interfaces

User interface design has become a substantial scientific and commercial endeavor in the past three decades (Carroll 1997). Much of this activity has been focused on graphical user interfaces, but a substantial body of knowledge has also been gathered for spoken interfaces (see e.g. (Cohen, Giangola, and Balogh 2004) for an overview). It is clear that much of this research is not directly applicable in the developing world, where technological sophistication and exposure to the commonplaces of the developed world cannot be assumed. Also, cultural factors – such as the prevalence of orally-dominated cultures – may have a significant impact on the design of interfaces in the developing world

(Goronzy et al. 2006). It is also expected that the preference for spoken interaction by users from the developed and developing world could be quite different, an aspect analysed by (Barnard, Plauché, and Davel 2008) with regard to the interplay between user sophistication and application complexity.

Research on user interface design in the developing world is nevertheless fairly rare; this statement is particularly true of spoken interfaces. This state of affairs is not surprising: the early-stage logistical challenges of preparing and designing persuasive user-interface experiments in the developing world are substantial. Therefore, only a small number of such experiments have been undertaken, and most initial studies (e.g. our own (Barnard, Cloete, and Patel 2003; Sharma et al. 2009)) have not produced conclusive evidence on interface design. This aspect of SDS design therefore becomes an increasingly important issue, and it is encouraging that research such as that reported in (Sherwani et al. 2009) points to real progress in understanding issues related to interface design.

In future research, we intend to devote an increasing amount of attention to the analysis of practical and theoretical issues related to user interfaces. Much of this research will take place in a series of pilot studies that are currently in progress, building on the infrastructure and expertise that have been gathered during our work to date.

Platforms and tools

In order to support telephone-based information access with speech technology, an integrated platform that connects to the public telephone network is required. Several commercial platforms with such functionality exist, but cost and licensing issues – which complicate the integration of new languages and capabilities – generally prevent their use in the developing world. Fortunately, there are a number of open-source initiatives that can be used as basis for the development of a telephony platform. Specifically, we have been working on the Asterisk platform (Digium-Inc. 2009), expanding it in a number of ways:

- ASR and TTS are supported through integration of ATK (a run-time speech-recognition front end) and Festival (the widely-used open-source TTS system) with Asterisk.
- We have developed an interface that makes the standard telephony functionality of Asterisk, along with the speech-technology enhancements, accessible through a Python programming interface.
- Tools have been written to support configuration, control and monitoring of the platform.

Taken together, these enhancements, which we call the “Lwazi platform”, make it possible for us to develop and deploy usable SDSs in a reasonably efficient manner. Of course, commercial platforms are more robust, better documented, and easier to use. However, our Asterisk-based platform is highly flexible, and freely available as open-source software.

In the same spirit, our group has developed and released a number of tools that support the creation of speech technologies in new languages. These include DictionaryMaker[1] (a toolkit that can be used for the efficient bootstrapping of pronunciation dictionaries), ASR-Builder[2] (tools used for training and experimenting with acoustic models for speech recognition) and Speect[3] (a modular toolkit for the development of TTS systems, designed to scale well with variable amounts of linguistic preprocessing available in a language). Each of these tools assists with a particular aspect of language-technology development, and all have been released as open-source software in the hope that others will find them useful and add to their functionality.

Selecting appropriate applications

Despite the strong need for enhanced information access in the developing world, it is not a trivial matter to select particular applications that are likely to have a sustainable positive impact. Even when all the technological hurdles have been overcome, a significant number of other variables need to be considered when planning such applications (Gumede and Plauché 2009; Plauché et al. 2010). Some of these variables are listed below.

- *Availability of information sources:* A useful information source must be reliable and current; thus, the back-end processes that are responsible for the maintenance of information that is provided by an SDS are crucial. Even in the developed world this poses significant integration problems between front-end and back-end systems, but in the developing world the relevant information sources must often be developed from scratch. Since the costs of the back-end systems often dwarf those of the SDS, the existence of reliable electronic information sources is a significant recommendation for a prospective application.
- *Financial sustainability:* It is a paradoxical fact that services (such as telephone call charges) are often highest in the developing world, where people earn the least. Services that provide useful information may therefore be unaffordable to their intended users unless special arrangements are made. In this regard sponsored lines or special arrangements with telecommunications operators are likely to be useful tools. Outbound calls or call-back options are also important tools in shielding end users from these costs.
- *Marketing to the target audience:* For users who have never used an Interactive Voice Response service or a Web browser, the concept of calling a telephone number not associated with a known contact person to obtain information can seem quite strange. Marketing the capabilities of an application and providing some basic education on the operation of such a service therefore become crucial to its success; locations where existing communities with shared interests gather (e.g. at health clinics or civic meetings) provide useful opportunities for such communications.

[1] http://dictionarymaker.sourceforge.net/

[2] http://asr-builder.sourceforge.net/

[3] http://speect.sourceforge.net/

In common with several other research groups (Plauché et al. 2006; Nasfors 2007; Sherwani et al. 2007; 2009), we have found that extensive interaction with potential target communities is critical to understand the practicalities of any intended application (Sharma et al. 2009; Gumede and Plauché 2009). It is likely that much will be learnt as increasing numbers of systems are piloted in the developing world, and that much of this learning will be transferable to widely dispersed regions – despite the inevitable local idiosyncrasies that appear in every pilot study.

Conclusion

The vision of large-scale information access in the developing world through spoken language systems remains ambitious, but substantial progress in that direction is being made. In tasks ranging from resource collection, through technology development and the design of suitable user interfaces, tangible outputs are appearing in several developing-world countries, suggesting that real impact may be on the horizon. Such impact may take different forms in different places: from health-care workers that use speech-enabled systems to manage their case loads, through rural farmers who obtain up-to-date technical and marketing information from such systems, to community members in remote locations who arrange transport and other logistics efficiently and reliably.

An important driver of technological progress is the sharing of insights across research groups with different perspectives, strengths, and challenges. (Such sharing was crucial – through initiatives such as the DARPA programs – to the rapid improvements culminating in market-ready speech technology in the developed world.) It seems as if there are many commonalities in developing-world speech technology, despite the superficial differences. Hence, it will be vital to develop strong cooperation across national boundaries to ensure that speech technology gains traction widely and expeditiously.

Although we have consciously focused on technical matters in the current contribution, it is worthwhile to mention the importance of financial models in this regard. Our own research has been very dependent on the availability of open-source components – not only toolkits such as HTK and Festival/Festvox, but also programming languages such as Python and the Linux operating system. For the extensive uptake of speech technology in the developing world, it seems important that this pool of available open technologies, and also open resources, must be expanded and strengthened wherever possible. It may seem attractive to make developing-world speech initiatives self-sustaining financially by restricting access to the technologies and resources developed for this purpose, and charging licensing fees for their use. However, we believe that such an arrangement is more likely to impede progress, since it restricts widespread access. Once these technologies have gained a sufficiently strong foothold, market forces will hopefully take over and enable sustainable business models – but relying on these forces prematurely will prevent the type of sharing that seems (to us) crucial at the current stage.

Looking ahead, one can envision speech technology as a significant force in bridging the digital divide. As cloud computation spreads its reach, it is increasingly likely that computer users in the developing world will use mobile telephones as the primary way to access information and computational resources. In such a scenario, spoken interfaces are an important way to overcome the interface limitations of a mobile handset – and therefore a significant factor in enabling the citizens of the developing world to participate in the information age.

Acknowledgment

The support for Project Lwazi by the Department of Arts and Culture of the Government of the Republic of South Africa is gratefully acknowledged. In particular, members of the Human Language Technologies Unit under Dr. Mbulelo Jokweni have been active partners and supporters of the efforts described here.

References

Barnard, E., and Wissing, D. 2008. Vowel variations in Southern Sotho: an acoustical investigation. *Southern African Linguistics and Applied Language Studies* 26:255–265.

Barnard, E.; Cloete, L.; and Patel, H. 2003. Language and technology literacy barriers to accessing government services. *Lecture Notes in Computer Science* 2739:37–42.

Barnard, E.; Davel, M.; and van Heerden, C. 2009. ASR corpus design for resource-scarce languages. In *Proceedings of Interspeech*, 2847–2850.

Barnard, E.; Plauché, M.; and Davel, M. 2008. The utility of spoken dialog systems. In *Proceedings of the IEEE Spoken Language Technology workshop*, 13–16.

Botha, G., and Barnard, E. 2005. Two approaches to gathering text corpora from the world wide web. In *Proceedings of PRASA*, 194.

Carroll, J. M. 1997. Human-computer interaction: Psychology as a science of design. *International Journal of Human-Computer Studies* 46:501 – 522.

Cohen, M.; Giangola, J.; and Balogh, J. 2004. *Voice User Interface Design*. Addison-Wesley.

Davel, M., and Barnard, E. 2008. Pronunciation predication with Default&Refine. *Computer Speech and Language* 22:374–393.

Davel, M., and Martirosian, O. 2009. Pronunciation dictionary development in resource-scarce environments. In *Proceedings of Interspeech*, 2851–2854.

Digium-Inc. 2009. Asterisk: The open source PBX & telephony platform. Online: http://www.asterisk.org.

Goronzy, S.; Tomokiyo, L.; Barnard, E.; and Davel, M. 2006. Other challenges: non-native speech, dialects, accents, and local interfaces. In Schultz, T., and Kirchhoff, K., eds., *Multilingual speech processing*. London: Academic Press. chapter 9, 273–317.

Gumede, T., and Plauché, M. 2009. Initial fieldwork for Lwazi a telephone-based spoken dialog system for rural

South Africa. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, 59–65.

Lehohla, P. 2003. *Census 2001: Census in brief*. Statistics South Africa.

Meraka-Institute. 2009. The Lwazi project. Online: http://www.meraka.org.za/lwazi.

Nasfors, P. 2007. Efficient voice information services for developing countries. Master's thesis, Department of Information Technology, Uppsala University.

Plauché, M.; Nallasamy, U.; Pal, J.; Wooters, C.; and Ramachandran, D. 2006. Speech recognition for illiterate access to information and technology. In *Proceedings of the IEEE International Conference on ICTD*, 83–92.

Plauché, M.; de Waal, A.; Sharma, A.; and Gumede, T. 2010. Morphological analysis: A method for selecting ICT applications in South African government service delivery. *Information Technologies and International Development (accepted for publication)*.

Sharma, A.; Plauché, M.; Kuun, C.; and Barnard, E. 2009. HIV health information access using spoken dialogue systems: Touchtone vs. speech. In *Proceedings of the IEEE International Conference on ICTD*, 95–107.

Sherwani, J.; Ali, N.; Mirza, S.; Fatma, A.; Memon, Y.; Karim, M.; Tongia, R.; and Rosenfeld, R. 2007. Healthline: Speech-based access to health information by low-literate users. In *Proceedings of the IEEE International Conference on ICTD*, 131–139.

Sherwani, J.; Palijo, S.; Mirza, S.; Ahmed, T.; Ali, N.; and Rosenfeld, R. 2009. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In *Proceedings of the IEEE International Conference on ICTD*, 447–457.

Tucker, R., and Shalonova, K. 2004. The Local Language Speech Technology Initiative. In *Proceedings of the SCALLA Conference*.

van Heerden, C.; Barnard, E.; and Davel, M. 2009. Basic speech recognition for spoken dialogues. In *Proceedings of Interspeech*, 3003–3006.

van Niekerk, D., and Barnard, E. 2009. Phonetic alignment for speech synthesis in under-resourced languages. In *Proceedings of Interspeech*, 880–883.

Zerbian, S., and Barnard, E. 2008a. Influences on tone in Sepedi, a Southern Bantu language. In *Proceedings of Interspeech*, 625.

Zerbian, S., and Barnard, E. 2008b. Phonetics of intonation in South African Bantu languages. *Southern African Linguistics and Applied Language Studies* 26(2):235–254.