UNEXPLORED DIRECTIONS IN SPOKEN LANGUAGE TECHNOLOGY FOR DEVELOPMENT

Frederick Weber[a] *Kalika Bali*[b] *Roni Rosenfeld*[c] *Kentaro Toyama*[b]

[a]Earth Institute, Columbia University, New York, USA
[b]Microsoft Research India, Bangalore, India
[c]Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

The full range of possibilities for spoken-language technologies (SLTs) to impact poor communities has been investigated only partially, despite what appears to be strong potential. Voice interfaces raise fewer barriers for the illiterate, require less training to use, and are a natural choice for applications on cell phones, which have far greater penetration in the developing world than PCs. At the same time, critical lessons of existing technology projects in development still apply and require careful attention. We suggest how to expand the view of SLT for development, and discuss how its potential can realistically be explored.

Index Terms**—** speech recognition, dialog systems, speech synthesis, ICT for development

1. INTRODUCTION

A wide range of information and communication technologies (ICTs) has been deployed in the service of development in Asia, Africa, and Latin America. In this paper, we discuss the related field of spoken language technologies for development (SLT4D), where technologies such as automatic speech recognition (ASR), dialogue systems, and text to speech (TTS) applications are applied to problems in the developing world [1]. The potential of SLT4D remains largely unexplored, despite offering a number of advantages in the context of development.

The most obvious advantage of SLT4D is in their use for illiterate or semi-literate populations, where text-based alternatives may not be applicable. Even well-schooled individuals can be illiterate in some of the languages they speak well, in multi-lingual countries like India, Ghana, or South Africa. SLT can even accommodate languages without any written form at all.

Another potential advantage of voice interfaces is that they might have shorter learning times than, for example, PC-based interfaces to the same task. We know that, at least in the developed world, most people are able to use interactive voice-response systems without any training, while use of a PC often requires significant exposure or formal training before practical use.

Third, in the developing world, a technology that presents a natural speech-based interface is penetrating more rapidly than any other technology, namely mobile phones. There are now more mobile phones than half of the population of the world [2]. In addition, mobiles offer a number of advantages, especially over more typical PC-based ICT4D deployments. Connectivity is much less susceptible to power outages, and covers a much wider area than internet penetration at much lower cost. Mobiles can be accessed in the field or at home, making it attractive for people on the go (doctors, NGO staff), the disabled, or applications with privacy needs. Finally mobile phones require less maintenance and servicing than PCs. It must be emphasized, though, that SLTs can have many non-telephone uses as well, and it isn't clear that the "killer app" of SLT4D will come from telephony.

The advantages discussed above apply to the end user as the beneficiary of SLT4D. However, SLTs can also be beneficial for organizations that work with the target populations. For example, given the prevalence of video, audio, and multimedia content in the service of development, SLTs could help index, search, and manipulate such data, with an NGO as the end user. This area of SLT4D has so far received no attention, yet may be one of the most promising.

2. SLT4D CAVEATS

If SLT holds significant potential in the context of the developing world, it also requires more resources and specific expertise than is typical of ICT4D, while sharing some of the same general constraints.

State-of-the art performance in speech recognition or text-to-speech requires the accumulation of significant data resources in the desired language. For ASR, these are principally annotated recordings from a variety of speakers (100 hours or more, covering the anticipated range of gender, accent, etc of the target population). For TTS, several hours of a single speaker are needed. For both, a pronunciation lexicon for the language is also needed. There is little commercial interest in developing these resources for the languages of developing countries, as state-of-the-art performance requires a greater financial investment that appears impossible to recoup from the target populations. However, there are several government and academic efforts underway to collect and codify language resources for underserved languages [3,4]. Also, for exploratory work much smaller datasets and time investments are sufficient [5], and accuracy below industry standards could be well-compensated for through careful application design. Still, without progress either on resources or the core speech technologies, full-scale systems of the quality of existing commercial applications in the developed world will be difficult to achieve.

The expertise required for creating speech technology for a new language is not common in the developed world, and is even rarer in developing countries. Tools are being built to simplify the task of developing speech technologies in new languages, with some measure of success, especially for TTS systems [6]. Unfortunately, specialists are still required to build effective speech applications currently. This human capacity must be expanded before SLT4D applications become widespread.

It should also be noted that speech as an interface modality has met with slow public acceptance in the developed world; it is more efficient to interact with either a human being or a computer screen than to conduct business over a strictly audio channel. Speech user interfaces (UIs), though arguably much simpler than the typical PC UI, still require some level of savvy, or experience with technology-driven interfaces. Especially among less educated populations, this barrier is likely to be higher. There is, for example, increasing evidence that with less formal education, hierarchical menus can be a challenge [25,10], or that users of automated systems don’t fully realize they are automated. This implies that the potential for SLT4D for a given application will depend on a “sweet spot” of target users in developing communities: those who “have the motivation and the skills to able to master the use of SLTs, yet for whom accessing ‘richer’ interfaces to information is not an option” [7]. The nature of such a “sweet spot” will vary by application and target population, but it some cases could be vanishingly small.

Finally, SLT4D projects are subject to the same requirements as any other technology-for-development effort. The intended beneficiaries are highly cost-sensitive, so affordability and a reasonable financial model are essential. Also, the information or service provided by a technological intervention can rarely effect change on its own – often physical infrastructure, organizational and political support, and human capacity are also necessary. Unless these aspects are simultaneously attended to, SLT4D projects will suffer the same fate as other ICT4D projects [24]. Thus, success will likely require sustained engagement with the end users and the organizations that work with them, leading to a longer development cycle with multiple design iterations.

3. RECENT EFFORTS

There have been only a handful of initiatives in SLT4D to date, mainly in the areas of dialog systems and TTS, although some other applications are being investigated.

One of the first dialog systems with a development goal was the Tamil Market (and subsequent Banana Crop SDS) pilot project [8,5]. These looked at providing weather, market prices, or agricultural extension information to rural farmers in Tamil Nadu, India. Both phone-based and kiosk-based multimodal interactions were considered. The project demonstrated that a speech recognition system in an underserved language could be built from very limited resources. Two projects currently underway are HealthLine [9,10], and Lwazi [11,12]. HealthLine targets community health workers, rather than the general population, and gives them speech-based access to a public health information database. This choice enables some level of training of the intended users, and allows for the continuous expansion and updating of the database they interrogate. Lwazi is a large-scale project currently in progress in South Africa, which aims to collect data in all 11 South African official languages, build core speech technologies, and implement development applications, with an emphasis on telephony.

Most recently, the Telecom Web (WWTW, aka “spoken web”) [13,14] project is seeking to develop and deploy an interconnected web of speech-driven “voice-sites” which are interoperable with the WWW, but intended to substitute for the WWW for illiterate or semi-literate individuals, including small business owners like autorickshaw drivers. Just like their text counterparts, the voice sites can support information access, information entry and transactions. A voice-driven generator of voice-sites has also been created and tested. Existing pilot studies include information and service delivery over the phone and a voice-site for grievance lodging.

Another area of effort is the development of text-to-speech systems in the languages spoken in the developing world. Several academic groups are already involved in generating the core TTS technology [15,16,17]. LLSTI [3] is a consortium created to coordinate these efforts and develop tools to open the field to less expert individuals or NGOs. Many of these groups are also constructing computer screen readers for the blind using the TTS systems they have developed.

Speech technology is also being applied to education, where it can supplement often-missing resources, be they human or textbook. An example of this is the use of speech recognition technology in literacy training: this has been developed in the US as Project LISTEN [18], but has recently been explored for English literacy in Ghana [19]. One way to provide textbook support is via an audio wiki, as tried in a pilot study in South African high school HIV AIDS education [20].

These exploratory projects give a sense of the potential of SLT4D, but much more needs to be done to move beyond the pilot stage and determine its true promise.

4. WHAT SHOULD BE DONE?

The field has so far been marked by mapping what has worked in the developed world to new contexts in the developing world. Though an appropriate initial strategy, this approach will meet with success, as Brewer *et al.* say, “accidentally at best” [21]. We believe there is broader research scope in SLT4D that extends far beyond re-targeting of existing SLT applications.

To that end, it is useful to consider various axes along which application designs can vary. If an application provides information, it can range from urgent (health emergency) to educational (audio wiki) to entertainment (sports scores). An application can target end users directly or through intermediaries, as HealthLine does. The form factors of the output—audio or video, cell phone or TV or PC monitor—can also be varied. So far entirely unexplored are uses of SLT4D “in the background” where end users do not necessarily interact directly with SLTs. We summarize below a set of possible research opportunities.

4.1 Fundamental Research

There are a number of basic research questions that still need to be addressed, in core speech technology, human-computer interaction (HCI), and sociology. Some of these apply to SLT applications in general but may require additional investigation in the context of illiteracy or diverse accents and dialects. For example, HCI questions such as interaction architectures, the correlation of information retention with recording length and complexity, the use of non-verbal audio cues, etc., are all relevant, and may require special investigation in the context of low-literate users.

For core speech technology, a range of questions arise. The variation in accent and dialect according to location, gender, ethnicity, and socioeconomic status is more pronounced in developing nations. This will require innovations in ASR model training and adaptation, as well as raise interesting questions regarding how to build TTS voices that are comprehensible and appropriate. Also, as we have mentioned, linguistic resources may be limited or missing altogether. Accommodating resource constraints will require innovations in building systems with limited or out-of-domain data, and in systems which evolve and adapt as they are used. Finally, simplifying and “democratizing” data collection and system building will encourage the spread of the technology into new languages and applications.

Equally important are sociological questions around technology diffusion in the target populations. Building cultural acceptance and trust in information or services provided by a new technology will be central to the success of any project [22,23]. An example is the emotive response to TTS or recorded real speech, which may differ from culture to culture, and could be an important concern for emotionally charged information such as healthcare. Understanding how existing technology, like cell phones, are already integrated into the social network [27] will be at least as important to successful outcomes as the technology itself.

4.2 Information Delivery

Most of the applications investigated so far are built around delivering information to a specific target group. These will be used according to their ability to fulfill compelling *needs* or *desires.* By *needs* we mean immediately useful information, such as health advice, actionable weather or market information, or educational information, such as retrieving training audio or video through a speech interface. *Desires* are primarily entertainment applications, from sports scores to music downloads. Though most existing applications are dialog systems that focus on *needs*, the best economic models might come in the *desires* category. As we mention above, information access applications are most vulnerable to the vanishing “sweet spot” when they target end users.

4.3 Task-driven applications

Perhaps more promising, then, are applications that accomplish a necessary task through speech technology. In this category we find literacy tutors like Project Listen or related language and pronunciation training. Document readers for the visually impaired, semi-literate, or second-language illiterate are also potentially compelling. Phrase-based translation systems could be useful in many contexts, from travel to interviews. Finally, dialog systems are likely to meet with easier acceptance as a means to an end, for example in mobile banking. Here speech simply broadens the reach of a system already in use.

4.4 Mediation

Among PC-based applications for development, there is increasing appreciation for human-mediated models of use, where a trained person mediates between the technology and the service client. Mediation has even proven useful in learning through video, where comprehension is not the issue [26]. Similarly, SLTs could be incorporated into larger human-mediated systems. For example, a human operator may issue a query that is then played back via TTS, or an IVR system may triage queries that ultimately go to human respondents.

4.5 Simpler SLT

In developing-world scenarios, simpler solutions are often best. Thus, before introducing an SLT, we should ensure there is no low-tech alternative. How far can application needs be met with recorded voice rather than TTS? Are there inexpensive ways to collect audio prompts needed even for projects with large vocabularies? What sorts of applications are there of audio databases that don't necessarily set out to recognize the spoken content?

4.6 Background SLT

A final line of possibilities include SLTs that are used in the background, without directly interacting with a user. For example, projects involving large video databases could apply speech-based word spotting as a way to search over the database without the need for human annotations and metadata. Similarly, opportunities may exist for non-profit organizations serving the poor. Could a speech interface facilitate health data collection on handheld PDAs carried by health workers?

5. CONCLUSIONS

Given the variety of roles SLTs could play in ICT4D, we believe there is much greater opportunity for exploring their use in development. As is always the case in ICT4D, however, care must be taken that projects consider the full context without narrowly focusing on the technology.

There remain significant challenges – and therefore, opportunities – for substantial research in SLT4D. There are fundamental research questions in HCI, core computational speech science, and the sociology of SLT acceptance and diffusion. Most existing SLT4D effort has been in information access; speech-enabled tasks, mediation, and "simple" or background uses of SLT are essentially untapped opportunities. With committed investment and careful design, the impact of SLTs in developing countries may prove to be far greater than in the developed world.

6. ACKNOWLEDGEMENTS

This paper grew out of discussions among the authors while F. Weber and R. Rosenfeld were visiting researchers at Microsoft Research Labs India. FW and RR are grateful to MSR India for its support and hospitality. RR is grateful to N. Balakrishnan for helpful discussions.

7. REFERENCES

[1] Sherwani, J and Rosenfeld, R.,, "The Case for Speech Technology for Developing Regions", in Proc HCI for Community and International Development, Florence, Italy, April, 2008.

[2] ITU statistical summary, http://www.itu.int/ITU-D/ICTEYE/Indicators/Indicators.aspx, accessed October 10, 2008.

[3] Tucker, R., and Shalonova, K., "The Local Language Speech Technology Initiative", in Proc. Crossing the Digital Divide—Shaping Technologies to Meet Human Needs, SCALLA Conference 2004, Kathmandu, Nepal, 2004

[4] van Rooyen, M., *et al., "*The Systematic Collection Of Speech Corpora For All Eleven Official South African Languages", in Proc. Spoken Language Technologies for Under-resourced Languages, Hanoi, Vietnam, May 2008

[5] Plauche, M., *et al.*, "How to Build a Spoken Dialogue System with Limited (or no) Resources". In Proc. AI in ICT for Development Workshop, Hyderabad, India, January, 2007.

[6] Black, A. & Schultz, T., "SPICE: Speech Processing—Interactive Creation and Evaluation Toolkit for New Languages", http://www.cmuspice.org

[7] Sherwani, J., "Are Spoken Dialog Systems Viable for Under-served Semi-literate Populations?", Ph.D. proposal, CMU, September 2006.

[8] Plauche, M., *et al.*, "Speech Recognition for Illiterate Access to Information and Technology". In Proc. Information & Communications Technologies and Development, Berkeley, USA, May 2006.

[9] Sherwani, J., *et al*., "HealthLine: Speech-based Access to Health Information by Low-literate Users". In Proc. Information & Communication Technologies for Development, Bangalore, India, December 2007.

[10] Sherwani, J., *et al.*, " Speech vs. Touch-tone: Telephony Interfaces for Information Access by Low Literate Users", Submitted for publication.

[11] Gumede, T, *et al.*, "Evaluating the Potential of Automated Telephony Systems in Rural Communities", submitted for publication, CSIR Annual Conference, 2008.

[12] Lwazi project website, http://www.meraka.org.za/lwazi/

[13] Agarwal, S, *et al.*, "Raising A Billion Voices", Interactions, March-April 2008.

[14] Kumar, A, *et al.*, "WWTW : The World Wide Telecom Web", Proc. *NSDR'07,* August 27, 2007, Kyoto, Japan.

[15] Tanuja Sarkar, *et al.*,, "Building Bengali Voice Using Festival", in Proc. of ICLSI 2005, Hyderabad, India, 2005.

[16] A G Ramakrishnan. *et al*., "VAACHAKA: A Kannada Text to speech synthesis system" in Proc. IEEE 8th International Symposium of Signal Processing and its Applications, 2005, Sydney, Australia

[17] J.A. Louw, *et al.*, "A general-purpose IsiZulu Speech Synthesiser", *South African Journal of African Languages*, 2006.

[18] Mostow, J., "Is ASR accurate enough for automated reading tutors, and how can we tell?" in Proc. Interspeech, Pittsburgh, USA, September 2006.

[19] Dias, B., *et al.*, "The TechBridgeWorld Initiative: Broadening Perspectives in Computing Technology Education and Research" in Proc. CWIT: Women and ICT, June 2005, Baltimore, MD

[20] Leinonen, T., *et al.*, "Audio Wiki for Mobile Communities: Information System for the Rest of Us", in Proc. Workshop on Speech in Mobile and Pervasive Environments, MobileHCI, Espoo, Finland, September, 2006.

[21] Brewer, E., *et al.*, "The Case for Technology for Developing Regions". IEEE Computer. Volume 38, Number 6, pp. 25-38, June 2005.

[22] Brand P and Schwittay A. "The Missing Piece: Human-Driven Design and Research in ICT and Development," in Proc. International Conference on Information and Communications Technologies and Development, Berkeley, USA, May 2006.

[23] Sandhu, J. and Agogino, A., "Design First, Technology Second", in Proc HCI for Community and International Development, Florence, Italy, April, 2008.

[24] Kuriyan, R. and K. Toyama, "Review of Research on Rural PC Kiosks", http://research.microsoft.com/research/tem/kiosks/Kiosks%20Research.doc, 2007.

[25] Grisedale, S., Graves, M and Grünsteidl, A. Designing a graphical user interface for healthcare workers in rural India, Proc. SIGCHI conference on Human factors in computing systems, Atlanta, USA, (1997), 471-478.

[26] Gandhi, R., R. Veeraraghavan, K. Toyama, V. Ramprasad. Digital Green: Participatory Video for Agricultural Extension, ICTD 2007, Bangalore, India. December 15-16, 2007.

[27] Donner, J., "The Use of Mobile Phones by Microentrepreneurs in Kigali, Rwanda: Changes to Social and Business Networks", *Information Technologies and International Development* 3 (2): 3-19.