Voice Search for Development

Etienne Barnard[1], *Johan Schalkwyk*[2], *Charl van Heerden*[1], *Pedro J. Moreno*[2]

[1]Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa
[2]Google Research, New York, NY, USA

ebarnard@csir.co.za, johans@google.com, cvheerden@csir.co.za, pedro@google.com

Abstract

In light of the serious problems with both illiteracy and information access in the developing world, there is a widespread belief that speech technology can play a significant role in improving the quality of life of developing-world citizens. We review the main reasons why this impact has not occurred to date, and propose that voice-search systems may be a useful tool in delivering on the original promise. The challenges that must be addressed to realize this vision are analyzed, and initial experimental results in developing voice search for two languages of South Africa (Zulu and Afrikaans) are summarized.

Index Terms: voice search, zulu, afrikaans

1. Introduction

In the developed world, systems utilizing speech technology have come a long way from the digit-recognition systems trialed at Bell Laboratories in 1952 to the current voice portals that are used by millions of callers per day. In fact, speech technology is now considered as a member of the exclusive club of Information Technologies with market sizes exceeding US$1 billion per year [1]; in countries such as the USA and Japan, speech technology is considered to be part of everyday life.

Speech technology has to date played a much smaller role in the developing world. Given the widespread belief that speech technology has particular potential in the developing world (where 98% of the illiterate or semi-literate people on earth live), and the rapid spread of telephone networks through the developing world, it is to be hoped that this situation will change significantly in the next decade.

This belief has encouraged a number of efforts to deploy speech-based systems in the developing world (see [2] for a recent review). Pilot systems using speech technology have been deployed in countries such as Kenya and Nepal, and there are even commercially sustainable systems in, for example, India and South Africa. Although these deployments have produced some encouraging results, it is uncontroversial that none of these systems have had a major impact within the communities where they were deployed.

The majority of these systems employ either kiosk-like interfaces, or spoken dialog systems based on directed dialogs. Both of these modalities mirror applications of speech technology in the developed world. Recently, a novel application of speech technology has attracted large-scale usage in the developed world, namely the use of speech recognition to perform searches through Web content and personal information[3]. This so-called voice search has a number of defining characteristics: the use of free-form input combined with massive language models, the ability to combine spoken and visual information on a mobile device, etc. In the current paper, we point out that voice search has the potential to deliver great impact in the developing world as well. In particular, we argue that some of the obstacles that have hampered the large-scale use of speech technology in the developing world can be overcome thourgh the use of voice-search applications.

Below, we review some of the challenges that must be overcome in order to develop high-impact speech systems in the developing world (Section 2), and present the reasoning behind our belief that voice search has a major role to play in overcoming these challenges. We investigate some of the practical issues that must be addressed if this vision is to become a reality in Section 3. Finally, we present initial results for two developing-world languages in Section 4.

Our examples are generally drawn from the sub-Saharan African context, since that is the developing-world region where we have carried out initial investigations. However, we believe that similar opportunities and challenges will occur throughout the developing world.

2. Developing-world challenges for speech technology

For speech technology to become a major factor in the developing world, progress on a number of scientific and technological fronts will be required. Fundamentally, it will be necessary to codify linguistic knowledge about the languages of the developing world in a way that supports technology development in those languages. During the past 40 years, the codification of such information in the languages of the developed world has become increasingly efficient and sophisticated: computer technologists and linguists have learnt how to extract and distil phonetic, phonological and syntactic information in a way that supports rapid and accurate technology development. Similar processes for the languages of the developing world will be required.

Fortunately, many of the methods and approaches that have evolved for the languages that already benefit from speech technology can be taken over directly for those of the developing world. For the choice of appropriate grapheme and phoneme sets, the development of pronunciation dictionaries, the collection of speech data and similar aspects of speech technology development, similar approaches are likely to be applicable in both circumstances. However, within these broad approaches there are details which are specific to certain languages (or language families) that may require solutions that were not relevant for languages with existing speech technology. For example, the earliest languages for which speech technology was developed were not tone languages; when speech technology for Mandarin was developed, innovations related to the modeling of pitch were therefore required. Similarly, it is possible that novel solutions will be required to deal with the click sounds that occur in some Southern Bantu languages, or the voicing

distinctions that act phonemically in several Southeast Asian languages.

It is clear that these technical challenges can be overcome with good scientific and engineering practice, as a number of successful speech-recognition and speech-synthesis systems in the developing world have demonstrated[2]. (Technical challenges certainly remain in obtaining sufficient efficiency in order to serve a significant fraction of the approximately 7,000 languages on earth). In pilot systems to date, the non-technical obstacles have, however, proven to be more troublesome. These include *economic*, *content-related* and *social* obstacles.

The economic obstacles stem from two sources. On the one hand, the collection of speech resources (such as sizeable corpora of speech recordings, which are required for high-quality speech recognition) is generally a major expense in typical developing-world environments. On the other hand, the cost of telephone calls in much of the developing world is too high (compared to typical incomes) to allow for experimental usage of voice services by local communities. (In India, mobile telephone calls are famously inexpensive, but this is not true in most developing countries.)

Digital content that is relevant to people of the developing world is generally scarce and distributed across numerous sources without any form of integration. Hence, developers of speech-based information systems face significant challenges in gathering, curating and presenting content that is truly useful to communities in the developing world.

Social obstacles to the adoption of new technologies (such as speech technology) have been researched extensively[2]. Non-industrialized communities tend to be conservative at least partially due to their limited exposure to rapid technological innovation. Hence, they are often seen as resistant to the uptake of such innovations.

Each of these obstacles can potentially be overcome within a voice-search framework. Voice search lends itself to efficient and low-cost data collection (thereby addressing resource constraints, and the data channel is typically one to two orders of magnitude less expensive than the voice channel for transmitting speech in the developing world. Voice search makes Web-based content available regardless of the original source of the data, which goes some way towards solving issues of content availability. Finally, voice search lends itself to technology appropriation: since it is a highly flexible approach to content access, local communities are able to adapt its usage to their own social contexts. (In this way, voice search builds on the flexibility of mobile-phone technology, which has indeed been appropriated in highly innovative ways e.g. through the use of missed calls[4] in the developing world.)

We therefore believe that voice search can play an important role in the adoption of speech technology in the developing world. Of course, this application introduces its own set of hurdles that must be overcome; some of which are discussed below.

3. Practical issues

To analyze the practical issues that must be addressed in pursuit of this vision, it is useful to distinguish between the two major algorithmic components of a voice-search system, namely the *acoustic model* and the *language model*. The technical issues can be broken down as follows:

- *How can accurate speech recognition be achieved* for users with limited technological exposure, when the amount of training data is relatively small compared to the amount of variability (in dialect, content, background conditions, etc.) within the usage domain? Broadly speaking, current state-of-the-art techniques solve this problem by (a) defining some basic linguistic resources, such as phoneme sets, (b) collecting a sufficiently varied speech corpus and (c) training an appropriately parameterized Hidden Markov Model.
- *How can language models be built* for languages in which prior resources (such as a large number of text-based Web queries) are not available? Again, relevant linguistic knowledge is required (e.g. for stemming or morphological analysis), appropriate corpora must be obtained or constructed, and a suitable finite-state language model is then derived.

The details of these issues are, of course, quite variable across the different languages of the developing world. We distinguish between three classes of languages, based on the amount of relevant resources available.

Languages such as Swahili and Afrikaans are well documented, with a rich literary tradition and extensive textual resources. These languages are nevertheless significantly deficient with respect to resources such as acoustic-phonetic information and speech corpora, and also with respect to the availability of speech technologists who are intimately familiar with the languages. The development of suitable acoustic models for languages in this class is relatively straightforward. The basic linguistic resources are easily gathered from the published literature or linguistic experts, and the available texts can be mined to create prompt sets that are used for the collection of corpora of read speech. (Such collection efforts benefit from the existence of large populations of literate mobile-phone owners speaking these languages, who can easily provide the contents of such corpora, utilizing purpose-made mobile-handset software for prompting and speech collection.) Language-model development is somewhat more challenging for these languages. For languages such as English, Mandarin Chinese and German, highly relevant language models can be constructed from large collections of text queries that have been collected for various purposes[3]. Such collections may not be available for even well-developed resource-constrained languages, and alternative approaches (e.g. language model construction from more general textual resources or language-identified Web content) will have to be investigated.

Many languages, such as Igbo, Zulu and Wolof, have been studied and documented extensively, but do not benefit from textual and linguistic resources comparable to those available in the mildly resource-constrained languages. Thus, the availability of suitable prompting material for speech corpus collection may be somewhat more problematic, and it may be necessary to translate such material from more resourced languages. By combining such translations with available text corpora and language-identified Web content, a diverse and (somewhat) relevant set of prompts for speech data collection can be created. Thereafter, the steps for acoustic-model training should be quite similar to those for the languages of the previous subsection: for sufficiently large languages, basic linguistic information, literate mobile-phone users and usable computer / telephone infrastructure will generally be available. Language-model development for these severely resource-constrained languages is likely to be a significant challenge because of the limited availability of suitable textual resources. Hence, translations of content from well-resourced languages are likely to be important in these instances. For most or all of these languages it is not realistic to assume the availability of machine-translation systems; hence, manual translations will be necessary. Manual transla-

tions are time-consuming and expensive activities for large corpora; hence, even if innovative approaches such as crowd sourcing are employed, language models constructed in this way are likely to cover only limited domains. Starting from these limited domains, a bootstrapping approach can be envisaged as a way to rapidly develop machine-translation capabilities for these languages. One ameliorating factor for language-model construction is the prevalence of proper names in typical Web queries. Such names are often unchanged across languages, so that relevant language-model content may be obtained from queries in well-resourced languages in the same geographical area as the target language. (Such languages exist for all three example languages mentioned above: English overlaps extensively with Igbo and Zulu, and French with Wolof). Since the amount of information available on the Web in these severely resource-constrained languages is likely to be quite limited, translation-based approaches may also be necessary to obtain useful *results* to queries in these languages. Statistical methods utilizing the translations employed for language-model construction may be useful in this regard, but (to our knowledge) there has to date been no research on this possibility.

There are numerous languages for which even the basic level of linguistic information assumed above is not available. Many of these languages are predominantly or exclusively spoken languages, so that the concept of text-based prompting does not make sense. For such languages, it does not seem feasible to consider voice-search solutions given the current state of technology (though some of the tools for data collection mentioned above may well prove useful for the process of documenting these languages).

A number of issues are likely to occur across a wide spectrum of languages in the developing world, regardless of the availability of resources. Some of these issues are discussed below.

- Many countries in the developing world are highly multilingual, with a world language such as English, French or Mandarin or regional language such as Swahili having a special status as the language of commerce or shared language. In those circumstances, it is likely that users are likely to switch between their local language and the shared language based on a complex set of factors. A useful system may therefore have to cater for two or more languages in parallel, and seamless operation in these languages without explicit choice by the user will be required. The shared language is likely to be strongly accented when used by the local population, and both acoustic and language models should therefore be developed for the particular target communities.
- Proper names play an important role in voice-search applications, and their pronunciations vary widely depending on factors such as (a) whether the user is more familiar with the spoken or written form of the name, (b) the users perception of the linguistic origin of the name, (c) the first language of the user and (d) the intended language produced by the user. High-accuracy voice search will require a detailed analysis of these factors and their relationships.
- Current voice-search implementations typically rely on visual presentations of search results: these are generally displayed on the screen of the users mobile device, for further manipulation by the user. This limits the usability of these services to users who have access to suitably feature-rich handsets and are sufficiently literate to process such information displays. In order to reach as many users as possible, it would be necessary to also consider users for whom these assumptions do not hold. (Given the rapid adoption of increasingly capable mobile handsets worldwide, and of tablet computing devices, the handset features may be less problematic in the near future.) This line of thinking prompts one to think of speech-only interfaces, with text-to-speech capabilities employed to present search results to the user. Such developments would require a new class of user interfaces, which would be somewhat similar to current spoken dialog systems.
- Similarly, most current voice-search systems rely heavily on data communications for several aspects of their operation (including the transmission of end-pointed and compressed speech). Many regions of the world currently have speech-only mobile telecommunications infrastructure, and would therefore require designs that use only the voice channel to support voice search. As with handset features, however, the rapid worldwide upgrading of network infrastructure may render this issue moot long before it can be considered seriously.

4. Initial results

In order to investigate several of the issues described above, we have developed initial voice-search systems for two of the under-resourced languages of South Africa, namely Zulu and Afrikaans. Zulu is a severely under-resourced language; a member of the Southern Bantu family, it is the home language of approximately 11 million people. Afrikaans is a Germanic language and home language of approximately 6 million people; we consider it so be mildly resource constrained.

Based on a random sample of text queries by anonymized computer users who had selected either of these two languages as their browser language, we estimated some statistics of query contents. In Table 2 we summarize the fraction of queries that were proper names, English words (English being a shared language in South Africa, according to the definition given in Section 3), and queries in the target language. It can be seen that usage profiles in these languages are quite different from one another: users who have selected Afrikaans as their browser language are far more likely to employ Afrikaans words in their queries. This probably results from the fact that there is a fair amount of Afrikaans content on the Web, whereas Zulu content is quite limited. We have therefore followed different strategies in developing our initial versions for the two systems. Whereas the Afrikaans system is primarily focused on Afrikaans content as well as proper names, our Zulu system combines a Zulu-and-proper-name system with a Zulu-accented English system (employing a language model constructed from localized English text queries).

Text queries drawn from collections obtained from the respective groups of users (those with Afrikaans browser settings for the Afrikaans system, and those with either English or Zulu browser settings for the Zulu system mixed in the ratio 55:45) were used as prompting materials for a data collection effort. (We also used a manually crafted finite-state grammar to generate an additional set of domain-specific Zulu prompts.) Data collection utilized software that was written for the Android telephone to (a) collect basic biographic information of the user, (b) prompt the user with a random sequence of the selected text queries, and (c) transmit the end-pointed and compressed spoken query to central servers[5]. Using a network of respondents, approximately 200,000 spoken queries were collected in each of

Table 1: *Voice Search results for Afrikaans and Zulu. System (1) is trained with ∼20k utterances, (2) with ∼90k and (3) with ∼180k utterances. The last row shows the results for our US baseline, which is currently the utimate target for all our languages.*

Exp.	NSACC		WSC1		WSC5		Ngauss		Nstates	
	Afr	Zul	Afr	Zul	Afr	Zul	Afr	Zul	Afr	Zul
ML (1)	56.8	54.1	58.5	55.0	62.1	56.8	8991	10040	493	560
Boosted MMI (1)	59.6	56.9	61.4	57.8	65.3	59.9	8991	10040	493	560
ML (2)	62.6	59.4	63.7	60.6	67.6	63.0	37005	39812	1857	1988
Boosted MMI (2)	65.7	63.8	66.6	65.1	70.7	67.5	37005	39812	1857	1988
ML (3)	63.5	62.0	64.6	63.3	68.5	65.7	69152	67855	3350	3250
Boosted MMI (3)	67.0	65.4	67.9	67.0	71.9	69.3	69152	67855	3350	3250
US baseline	72.3		74.8		79.0		327016		7959	

the target languages. Language models were constructed from

Table 2: *Distribution of text queries, by chosen browser language.*

Chosen browser language	English	Proper names	Browser language	Mixed & other
Zulu	33%	38%	9%	20%
Afrikaans	9%	40%	49%	2%

these same text queries (with certain queries held out to serve as test set), and acoustic models were trained from the prompted spoken queries (again, with certain speakers data held out for testing purposes). We did not perform manual transcription or verification of the queries, simply using the prompted text as transcriptions during the training process. Perplexity was measured on the held out test set to evaluate the quality of our language models as shown in table 3. Our initial acoustic models

Table 3: *Language model statistics and evaluation results for Afrikaans and Zulu.*

	Vocab size	# ngrams	PPL	OOV
Afr	775907	10438684	113.95	0.057%
Zul	77038	1939227	76.55	0.061%

are standard 3-state context-dependent (triphone) models with variable numbers of Gaussians per state. These were trained on 39-dimensional PLP cepstral coefficients, with CMN, LDA and STC transforms applied. In table 1, we show recognition results on the same held-out test set we used to measure perplexity. Results are presented after ML and boosted MMI objective function optimization, and are shown for systems trained on variable amounts of data to give comparative results to typical low-resource language corpora. We used three measures to analyze the quality of these systems. We measure the normalized sentence accuracy (NSACC) to account for mismatches based on plural "s", apostrophes, hyphens and white spaces. For WSC1 (web score top 1) a we consider a query correct if top search result in a search engine is the same for the transcription truth and the recognition result. For WSC5, we look for at least one common result among the top 5 search results.

Table 1 indicates that our systems have error rates 20% to 50% higher (relatively) than the state-of-the-art US English system, by all three measures. This is encouraging given the extensive user acceptance of that system. The small differences between systems (2) and (3) – trained with 90,000 and 180,000 utterances, respectively – suggests that the smaller amount of training data may be sufficient for initial releases (prior to refinement with field data, which is likely to be crucial for novel customer-facing speech systems).

5. Conclusion

We have provided a number of arguments suggesting that voice search may play an important role in increasing the impact of speech technology in the developing world. These include economic considerations, factors related to content availability, and social factors.

Of course, voice search requires a speaker-independent speech-recognition system with a large vocabulary, high perplexity and complex language model; developing such a system for developing-world applications runs counter to the prevalent wisdom on how speech technology should be introduced into such environments. In addition, developing-world applications of voice search will introduce a whole set of additional challenges, some of which are discussed in Section 3.

We nevertheless believe that the potential benefits from operational voice-search systems in the developing world are tremendous. In developed countries, information technology has had a tremendous impact during the past five decades, and voice search promises to deliver some of the capabilities of the information age to developing-world citizens – with a reach that is not foreseeable with any other current technology. Hence, the effort required to overcome these obstacles is well warranted. Such effort will also have spin-offs in the development of speech resources, modules and applications for several languages that are currently not supported by speech technology.

Our initial experiments with two languages of South Africa have shown promising results; we intend to explore their usage in much more detail during coming years, and also to extend our development to many more languages of the developing world.

6. References

[1] E. Benhamou, J. Eisenberg, and R. H. Katz, "Assessing the IT R&D ecosystem," *Communications of the ACM*, vol. 53, no. 2, pp. 76 – 83, 2010.

[2] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh, "Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india," in *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*. New York, NY, USA: ACM, 2010, pp. 733–742.

[3] B. Erol, J. Cohen, M. Etoh, H.-W. Hon, J. Luo, and J. Schalkwyk, "Mobile media search," in *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 4897–4900.

[4] J. Donner, "The rules of beeping: Exchanging messages via intentional "missed calls"," *Journal of Computer-Mediated Communication*, vol. 1, no. 13, 2007.

[5] T. Hughes, K. Nakajima, L. Ha, P. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," Makuhari, Japan, September 2010, (Accepted for publication at Interspeech 2010).