Voice as Data: Learning from What People Say

Tapan S. Parikh
UC Berkeley School of Information
parikh@berkeley.edu

Introduction

Development is fundamentally about understanding people, their Motivations, behaviors and reactions. We have two primary means of understanding people — observing what they do, and what they say. As the AI4D community has noted, people's increased use of mobile devices has led to a wealth of new data relevant to these topics. We are on the cusp of developing incredibly powerful tools that can help us understand how human beings migrate, transact and acquire wealth. This could have a large impact on how we determine policies and allocate resources.

Most of this analysis has tended to focus on what people *do* — where they go, who they talk to, what they buy, etc. (Eagle and (Sandy) Pentland 2006). I argue that what people *say* is an equally rich source of development data, often containing information that cannot be obtained from people's actions, such as their needs, hopes and aspirations. Voice is the most natural form of communication, especially for people who speak a non-mainstream language, and/or have marginal literacy skills. These are often exactly those populations who are most disenfranchised, and therefore most need their voices to be heard.

Mobile phones are a natural channel for improving voice communications between development institutions and the communities they serve. However, voice communications is traditionally 1-to-1, or 1-to-many, but not many-to-many. In general, it is still difficult to process and present large amounts of heterogeneous voice content. There are several challenges to this. First, there are obvious privacy concerns. Voice content is also inherently difficult to aggregate, structure, search or scan. In this paper, I describe how my group is exploring the use of voice as a public knowledge medium. I also discuss some approaches to managing large amounts of voice content, leveraging a combination of user feedback, collective human intelligence and intelligent automation.

Voice as a Knowledge Medium

In this section, I describe two applications using voice as a medium for aggregating and communicating knowledge.

Voice-based Social Media

Uneducated and/or illiterate people are used to accessing and communicating information through oral means. Oral communication is two-way, involving a real audience, as opposed to writing, for which the audience can be abstract, temporally and spatially removed, or not exist at all. According to prior empirical research, oral thinking is *situational* — bound to a specific time, place and set of needs (Ong 2002). It is also *aggregative* — adding to a potentially conflicting and/or redundant knowledge base, rather then refining a self-consistent set of definitions.

Social media such as mailing lists, forums or message boards are particularly well-suited for these affinities. In collaboration with IBM Research India and the Development Support Centre (a local NGO), we have designed, implemented and deployed *Avaaj Otalo* ("voice porch"), the voice-based equivalent of an online discussion board, for small farmers in Gujarat, India (Patel and others). Farmers can record their questions using a toll-free number, which are replied to by other farmers, or by experts working for the NGO. Farmers can also browse prior questions and answers, finding responses to their queries, or for general learning. The most popular questions and answers are re-broadcast on a local agricultural radio program, for wider dissemination.

Avaaj Otalo provides farmers with a knowledge base that is accessible, relevant and credible. This system has been deployed on a pilot basis since January 2009, with access provided to fifty incoming phone numbers strategically distributed throughout the state. The system has averaged over 1000 calls per month, with many of the answers being provided by other farmers (as opposed to experts, reducing a significant information bottleneck). One farmer self-reported an increase in income of over $3,000 due to information he obtained through the system. The farmers' questions themselves are a rich source of knowledge for understanding the needs and issues facing small farmers, and their answers represent an equally rich source of local innovations and agricultural practices, both of which could be used as inspiration for new policies, projects or initiatives aimed at small farmers.

Voice-based Data Collection

Collecting accurate and timely information from rural communities is important for many purposes — including monitoring projects, conducting clinical trials, responding to emergencies, maintaining medical records, managing supply chains, providing financial services, doing epidemiological surveillance and for conducting a survey or census. Many groups are investigating the use of mobile phones for collecting data. A recent study has shown that collecting data using voice can significantly reduce errors, when compared to using interactive forms or composing a structured text message (Patnaik and others).

(Patnaik and others) relied on a call center approach, where respondents called dedicated staff who asked them questions and confirmed and recorded responses. This incurs significant overhead in terms of phone airtime and centralized office infrastructure and human resources. One alternative that we are investigating is to allow field staff to capture audio transcripts of their field interviews, which can later be annotated and/or transcribed by trained data entry staff in batch mode. Separating data collection from entry could reduce airtime usage (because data could be transferred asynchronously), and allow for less staff to handle the actual entry in batch mode.

Audio is also an effective medium for self-reporting, particularly of sensitive data. Audio computer-assisted self-interviewing, or its telephonic variant, T-ACASI, is used by psychologists, clinicians, social scientists and development practitioners as a cheap and private way of collecting private, anonymous information from subjects (Cooley and others 2000). Deploying such systems, accessible to remote populations using a low-cost mobile phone, could drastically reduce the costs and hence increase the reach of data collection. Timely and accurate responses would no longer depend on a trained fieldworker being present at exactly the right time and place.

Challenges and Opportunities

In this section, I outline two challenges to (and consequent opportunities for) improving the functionality and usability of voice knowledge systems.

Voice Search

Avaaj Otalo offers very limited navigation features. Farmers (and experts) must listen to all previous questions and answers before they can find those relevant to them. Even a rudimentary interface for tagging and/or searching could dramatically improve the efficiency of this system. Indexing and search of complete voice transcripts is likely to be infeasible for low-resource languages, at least with current speech recognition technologies (see below). We are exploring several alternative ways to address this. First, as search is already a "noisy", potentially error-prone task, we can use relatively low-grade speech recognition for clustering and/or categorizing highly related content. Second, we can use contextual factors obtained from users (including location, language, dialect and inferred social and economic characteristics) to further group content. Third, we can explicitly *crowd-source* tags for voice content, either by users as part of the applications themselves, or using games, micro-payments and other incentives. Finally, we can leverage feedback from users (for example, "click-through" rates, or the equivalent), to improve the ranking of popular results.

Speech Recognition

Collecting speech training data is tedious, expensive and labor-intensive (van Rooyen et al.). Voice-based knowledge systems, such as those described earlier, can generate a wealth of audio data for training purposes. This data must be annotated and/or transcribed before it can be used for training a speech recognizer or other natural language system. We can leverage increased access to mobile phones, and a distributed network of native speakers, to address this problem. An automated system could call and replay important terms to users who could transcribe them using a reply SMS message, or just repeat them over the phone (in the process, generating additional pronunciations) (Ledlie and others). Repeating utterances could also be a form of voice CAPTCHA that a user must perform before using an application or service. By generating transcriptions from multiple sources, and potentially for multiple versions of the utterance, we can aim for maximum coverage. Because the transcription task is verifiable, it is more likely to generate correct responses. We can also tie incentives to matching consensus transcriptions. Payments can be issued in terms of airtime, further reducing transaction costs. Transcription can also done as an educational activity for students, for example for learning the English alphabet.

Conclusion

Making voice a first-class data type could dramatically increase public participation in knowledge creation. I have described two applications and related technologies that explore this idea, and could make locally generated voice content more amenable to large-scale access and analysis.

References

Cooley, P. C., et al. 2000. Automating telephone surveys: Using T-ACASI to obtain data on sensitive topics. *Computers in Human Behavior* 16(1):1 – 11.

Eagle, N., and (Sandy) Pentland, A. 2006. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.* 10(4):255–268.

Ledlie, J., et al. Crowd translator: On building localized speech recognizers through micropayments. In *NSDR 2009.*

Ong, W. J. 2002. *Orality and Literacy: The Technologizing of the Word.* Routledge.

Patel, N., et al. Avaaj Otalo - A field study of an interactive voice forum for small farmers in rural India. In *CHI 2010.*

Patnaik, S., et al. Evaluating the accuracy of data collection on mobile phones: A study of forms, SMS, and voice. In *ICTD 2009.*

van Rooyen et al., M. The systematic collection of speech corpora for all eleven official South African languages. In *SLTU 2008.*