Speech and Multimodal Interaction in Mobile Search

Junlan Feng
AT&T Labs - Research
180 Park Ave
Florham Park, NJ USA
junlan@research.att.com

Michael Johnston
AT&T Labs - Research
180 Park Ave
Florham Park, NJ USA
johnston@research.att.com

Srinivas Bangalore
AT&T Labs - Research
180 Park Ave
Florham Park, NJ USA
srini@research.att.com

ABSTRACT

This tutorial highlights the characteristics of mobile search comparing with its desktop counterpart, reviews the state of art technologies of speech-based mobile search, and presents opportunities for exploiting multimodal interaction to optimize the efficiency of mobile search. It is suitable for students, researchers and practitioners working in the areas of spoken language processing, multimodal and search with an emphasis on a synergistic integration of these technologies for applications on mobile devices. We will provide detailed bibliography and sufficient literature for everyone interested to jumpstart work on this topic.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Design

Keywords

Mobile voice search, Speech recognition, Multimodal Interface

1. BACKGROUND

With the deep penetration of high-speed wireless networks symbiotically complemented by the burgeoning demand for smart mobile devices, mobile devices are expected to overtake personal computers (PCs) as the most popular way to access the Internet in the near future. Mobile search applications such as web search, local listing search, and question answering are some of the leading downloaded applications contributing to this trend.

However, the small screen and small keyboard of most mobile devices currently limits the effectiveness of these applications. At the same time, certain mobile devices are endowed with interface capabilities that are beyond those of a typical desktop computer, such as touch screens, device movement sensors, cameras, and built-in microphones. These functionalities provide new opportunities and challenges for mobile application design. Multimodal interfaces combining spoken interaction with dynamic information content have the potential to make mobile devices more favored for information search and access. In this tutorial, we address these issues in great detail.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. SCOPE OF THE TUTORIAL

This tutorial covers challenges, typical applications, technologies, and multimodal interfaces. In particular, we will address the following topics.

2.1 Comparison of Mobile search and desktop search

Mobile Web search is inherently different than its desktop counterpart. In the tutorial, we will highlight the differences from the following perspectives.

1. Richer interaction context: A mobile device is out and about in the open world with the bearer of the device. This aspect brings several advantages and challenges to mobile search. Situated in a rich real world context, a mobile search application has access to latitude/longitude coordinates of the device which can be exploited to deliver location-aware search results in a location-sensitive user interface.

2. Browsing/Search behavior: Users' information needs are more precise and immediate in the mobile medium. While users on a PC utilize browsing (exploratory queries) and searching (targeted queries) behavior to an equal extent, users on a mobile device are more likely to search for targeted information like weather, movies, and nearby restaurants. Furthermore, beyond searching for information, users' are more likely to complete their task by following the search results (e.g. reserve a restaurant or purchase a ticket), due to the convenience provided by the mobile device.

3. Personalization: Mobile devices are often a personal tool, and so all major components of a voice search system can be made to behave robustly and cooperatively for the specific user of that device

4. User Interface: Given the limited screen real-estate, the interface capabilities of mobile devices are very different from a typical PC and hence the interaction needs to be suitably tailored to the mobile device.

2.2 Landscape of mobile voice search applications

In this part of the tutorial, we will highlight through examples, various voice-enabled mobile search applications that

have been built in the recent few years. We will segment these applications into three generations.

1. First generation: Early voice search mobile applications focused on directory assistance. The most traditional applications are commercial 411 business listing services, which were implemented as a voice-only two exchange dialog such as 800-FREE-411, 800-CALL-411, and GOOGLE-411.
2. Second generation: Local search, an extension of business directory search, is one of the leading trends of search applications on mobile phones. Most recent local search applications target smart phones and offer searches of business listings, maps, directions, movie schedules, local events, travel resources such as airlines, hotels and rental cars. Examples of mobile local search applications include YPMobile, Speak4it, and Bing Local voice search. Voice-enabled web search on mobile phones is now available and free from major search engine providers.
3. Third generation: Voice-enabled question answering is a relatively new direction for voice search. The ChaCha iPhone application, developed using AT&T's speech recognition, allows people to ask any question in natural language. Applications such as Vlingo and Siri go beyond search to provide capabilities such as messaging and fulfillment of ticket bookings, reservations, etc.

2.3 Automatic speech recognition in mobile context

Having motivated the need for a new and different user interface to support interaction on a mobile device, in this part of the tutorial, we will present, in detail, the components and algorithms of a speech recognition system.

Automatic Speech Recognition (ASR) is a task of finding the most likely set of words for a given acoustic signal. There are three key models in ASR, namely, pronunciation models, acoustic models and language models. In the past three decades, there have been tremendous technology progresses towards higher ASR accuracy, speed, and robustness. In this tutorial, we will summarize recent advances in ASR that are relevant specifically for mobile voice search.

An acoustic model (AM) is created by taking audio recordings of speech and their transcription as input and compiling them into a statistical representations of the sounds that make up each word. State-of-the-art ASR systems use Hidden Markov Models (HMMs) for the acoustic model. The challenge is the presence of ambient noise in mobile environment significantly impacts ASR accuracy. However, there are some unique opportunities as well when using mobile devices. Due to the personal nature of these devices combined with the locations from where they are being used from, general purpose acoustic models can be adapted to a specific speaker and context to optimize ASR performance. We will overview different techniques for both speaker adaptation and location adaptation.

Language models(LM) represent the probability of the sequence of words. LM in voice search often have very high perplexity and large vocabulary sizes. We will present a few directions that researchers and engineers have been recently exploring to improve the quality of the language model and constrain the language models.

Pronunciation models(PM) translate words into phonemes. Most large vocabulary speech recognition systems make use of a dictionary with multiple possible pronunciation variants per word, which are either provided by trained developers or generated from letter-to-phoneme (L2P) rules. However, in mobile search tasks, many query terms and targeted information fields include entities such as names of people, businesses, cities, movies, music, etc., of which the variations in pronunciation among different individuals have been a big challenge for ASR. We will overview a few approaches proposed in recent years to address this challenge. We will categorize them into two categories: knowledge-based approaches and data-driven approaches.

2.4 Robust Integration of Speech Recognition and Search

In this part of the tutorial, we will address the robustness issue of voice search given that speech recognition is an error-prone process. Voice search is essentially an integration of automatic speech recognition (ASR) and text or database search. Researchers have proposed various approaches to tightly coupling the two components for better overall system performance. There are also research efforts devoted to re-design search specifically for voice queries.

2.5 Contextually-situated search

Search on mobile devices is inherently situated, it takes place on a particular device in a particular location at a particular point in time. In order to behave robustly and cooperatively, mobile search systems need to incorporate contextual information through all levels of processing from speech recognition, through query understanding, through search and user interface design. Contextual factors include, the spatial location of the user, the time at which the user submits the query, the velocity at which the user is traveling and the social network the user is connected to. In this tutorial, we will address the challenges posed by each kind of contextual information and technical approaches to incorporating into mobile search.

2.6 The role of multimodal interfaces in enhancing voice enabled mobile search

One of the key differentiators between contemporary mobile search systems and more traditional voice activated information systems is their utilization of a multimodal interface where in addition to voice input and output, the user is also presented with a graphical display.

In this tutorial, we will present both the technical challenges of this more freeform style of interaction and the many advantages of having a persistent graphical display for information presentation and user input as listed below.