**Voice recognition** - ***Google Search by Voice***

Introduction:

Using our voice to access information has been part of science fiction ever since the days of Captain Kirk talking to the Star Trek computer. Today, with powerful smartphones and cloud based computing, science fiction is becoming reality. In this chapter we give an overview of Google Search by Voice and our efforts to make speech input on mobile devices truly ubiquitous.

The explosion in recent years of mobile devices, especially web-enabled smartphones, has resulted in new user expectations and needs. Some of these new expectations are about the nature of the services - e.g., new types of up-to-the-minute information (”where’s the closest parking spot?”) or communications (e.g., ”update my facebook status to ’seeking chocolate’”). There is also the growing expectation of ubiquitous availability. Users increasingly expect to have constant access to the information and services of the web. Given the nature of delivery devices (e.g., fit in your pocket or in your ear) and the increased range of usage scenarios (while driving, biking, walking down the street), speech technology has taken on new importance in accommodating user needs for ubiquitous mobile access - any time, any place, any usage scenario, as part of any type of activity.

A goal at Google is to make spoken access ubiquitously available. We would like to let the user choose - they should be able to take it for granted that spoken interaction is always an option. Achieving ubiquity requires two things: availability (i.e., built into every possible interaction where speech input or output can make sense), and performance (i.e., works so well that the modality adds no friction to the interaction).

This chapter is a case study of the development of Google Search by 1 Voice - a step toward our long term vision of ubiquitous access. While the integration of speech input into Google search is a significant step toward more ubiquitous access, it posed many problems in terms of the performance of core speech technologies and the design of effective user interfaces. Work is ongoing - the problems are far from solved. However, we have, at least, achieved a level of performance such that usage is growing rapidly, and many users become repeat users.

Google Search by Voice:

Mobile web search is a rapidly growing area of interest. Internet-enabled smartphones account for an increasing share of the mobile devices sold throughout the world, and most models offer a web browsing experience that rivals desktop computers in display quality. Users are increasingly turning to their mobile devices when doing web searches, driving efforts to enhance the usability of web search on these devices. 5 Although mobile device usability has improved, typing search queries can still be cumbersome, error-prone, and even dangerous in some usage scenarios.

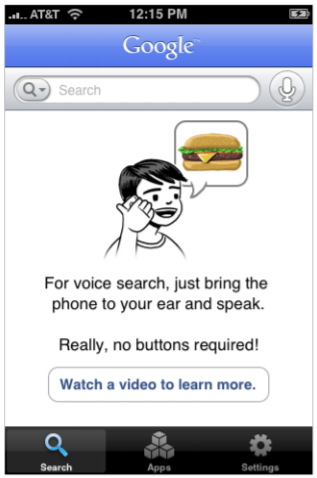


Figure 4: Google search by Voice for iPhone.

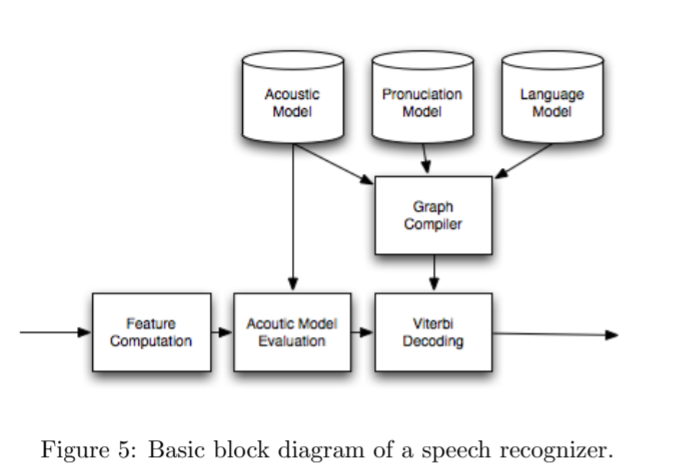
In November 2008 we introduced Google Mobile App (GMA) for iPhone (Figure 4) that included a search by voice feature. GMA search by voice extended the paradigm of multi-modal voice search from searching for businesses on maps to searching the entire world wide web. In the next few sections we discuss the technology behind these efforts and some lessons we have learned by analyzing data from our users.

Technology:

The goal of Google search by Voice is to recognize any spoken search query. Table 1 lists some example queries, hinting at the great diversity of inputs we must accommodate. Unlike GOOG-411, which is very domain-dependent, Google search by Voice must be capable of handling anything that Google search can handle. This makes it a considerably more challenging recognition problem, because the vocabulary and complexity of the queries is so large (more on this later in the language modeling Section

**Example Query**: Example queries to Google search by Voice.

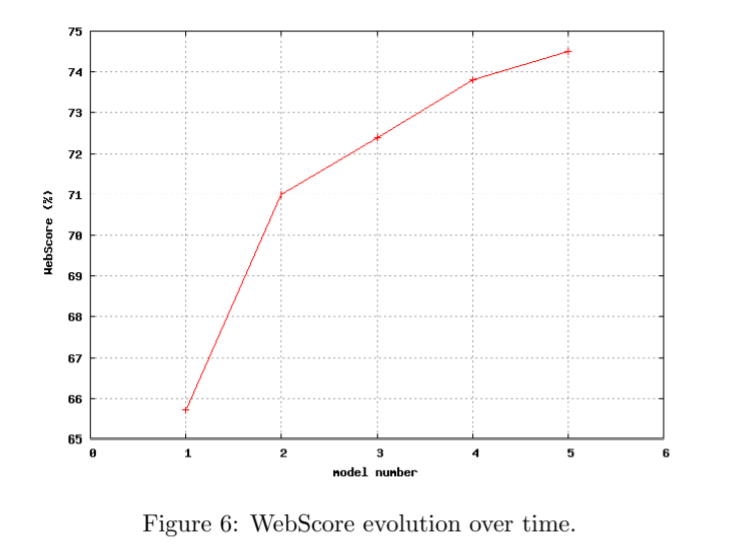
* images of the grand canyon
* what’s the average weight of a rhinoceros
* map of san francisco what time is it in bangalore
* weather scarsdale new York
* bank of america dot com
* A T and T



Google search by Voice. For each key area of acoustic modeling and language modeling we will describe some of the challenges we faced as well as some of the solutions we have developed to address those unique challenges. In Section 3.1 we will review some of the common metrics we use to evaluate the quality of the recognizer. In Sections ,we describe the algorithms and technologies used to build the recognizer for Google search by Voice.

Accuracy of an Evolving System:

The basic form of the acoustic models used are common in the literature. The experiments shown here all use 39-dimensional PLP-cepstral [5] coefficients together with online cepstral normalization, LDA (stacking 9 frames), and STC [3]. The acoustic models are triphone systems grown from decision trees, and use GMMs with variable numbers of Gaussians per acoustic state. We optimize ML, MMI, and ’boosted’-MMI [8] objective functions in training. Figure 6 shows the accuracy of the system on an off-line test set across various acoustic models developed in the first year of production. Each point on the x-axis represents a different acoustic model. These evaluations all use the same production language model (LM) estimated toward the end of the first year of deployment, but change the underlying acoustic model. The test set has 14K utterances and 46K words. The metric used here is WebScore, described above, which provides a measure of sentence-level semantic accuracy.



The first point on the graph shows the baseline performance of the system with mismatched GOOG-411 acoustic models. The second point, model 2, largely shows the impact of matching the acoustic models to the task using around 1K hours of transcribed data. For model 3, we doubled the training data and changed our models to use a variable number of Gaussians for each state. Model 4 includes boosted-MMI and adds around 5K hours of unsupervised data. Model 5 includes more supervised and unsupervised data, but this time sampled at 16KHz.

Next Challenges:

The growing user base of voice search together with Google’s computational infrastructure provides a great opportunity to scale our acoustic models. The inter-related challenges include how and where to add acoustic parameters, what objective functions to optimize during training, how to find the optimal acoustic modeling size for a given amount of data, how to field a realtime service with increasingly large acoustic models, and how to get reliable labels for exponentially increasing amounts of data. Early experiments in these directions suggest that the optimal model size is linked to the objective function: the best MMI models may come from ML models that are smaller than the best ML models; that MMI objective functions may scale well with increasing unsupervised data; that speaker clustering techniques may show promise for exploiting increasing amounts of data; and that combinations of multi-core decoding, optimizations of Gaussian selection in acoustic scoring, and multi-pass recognition provide suitable paths for increasing the scale of acoustic models in realtime systems.