Speech Recognition for Mobile Devices at Google

Mike Schuster

Google Research, 1600 Amphitheatre Pkwy., Mountain View, CA 94043, USA
schuster@google.com

Abstract. We briefly describe here some of the content of a talk to be given at the conference.

1 Introduction

At Google, we focus on making information universally accessible through many channels, including through spoken input. Since the speech group started in 2005 we have developed several successful speech recognition services for the US and for some other countries. In 2006 we launched GOOG-411 in the US, a speech recognition driven directory assistance service which works from any phone. As smartphones like the iPhone, BlackBerry, Nokia s60 platform and phones running the Android operating system like the Nexus One and others becoming more widely used we shifted our efforts to provide speech input for the search engine (Search by Voice) and other applications on these phones. Many recent smartphones have only soft keyboards which can be difficult to type on, especially for longer input words and sentences. Some Asian languages, for example Japanese and Chinese are more difficult to type as the basic number of characters is very high compared to Latin alphabet languages. Spoken input is a natural choice to improve on many of these problems, and more details are discussed in the sections below.

We have also been working on voice mail transcription and YouTube transcription for US English, which are also publically available products in the US, but the focus here will be on speech recognition in the context of mobile devices.

2 GOOG-411

GOOG-411 is Google's speech recognition based directory assistance service operating in the US and Canada [1], [2]. This application uses a toll-free number, 1-800-GOOG-411 (1-800-4664-411). The user is prompted to say city, state and the name of the business s(he) is looking for. Using text-to-speech the service can give address and phone number, or can connect the user directly to the business. As backend information from Google Maps Local is used.

While this is a useful application to search for restaurants, stores etc. it is limited to businesses. Other difficulties with this kind of service include the

B.-T. Zhang and M.A. Orgun (Eds.): PRICAI 2010, LNAI 6230, pp. 8–10, 2010.

necessity of a dialog, relatively expensive operating costs, listing errors in the backend database, and most importantly to not be able to give richer information (as on a smartphone screen) back to the user.

3 Voice Search

In 2008 Google launched Voice Search in the US for several types of smartphones [3]. Voice Search adds simply the ability to speak a search query to the phone instead of having to type it into the browser. The audio is sent to Google servers where it is recognized and the recognition result along with the search result is sent back to the phone. The data goes over the data channel instead of the voice channel which allows higher quality audio transmission and therefore better recognition rates. Our speech recognition technology is relatively standard, below some details.

Front-End and Acoustic Model. For the front-end we use 39-dimensional PLP features with LDA. The acoustic models are ML and MMI trained, triphone decision-tree tied 3-state HMMs with currently up to 10k states total. The state distributions are modeled by 50-300k diagonal covariance Gaussians with STC. We use a time-synchronous finite-state transducer (FST) decoder with Gaussian selection for speedy likelihood calculation.

Dictionary. Our phone set contains between 30 and 100 phones depending on the language. We use between 200k and 1.5M words in the dictionary, which are automatically extracted from the web-based query stream. The pronunciations for these words are mostly generated by an automatic system with special treatment for numbers, abbreviations and other exceptions.

Language Model. As our goal is to recognize search queries we mine our language model data from web-based anonymous search queries. We mostly use 3-grams or 5-grams with Katz backoff trained on months or years of query data. The language models have to be pruned appropriately such that the final decoder graphs fit into memory of the servers.

Acoustic Data. To train an initial system we collect roughly 250k of spoken queries using an Android application specifically designed for this purpose [4]. Several hundred speakers read queries off a screen and the corresponding voice samples are recorded. As most queries are spoken without errors we don't have to manually transcribe these queries.

Metrics. We want to optimize user experience. Traditionally speech recognition systems focus on minimizing word error rate. This is also a useful measure for us, but better is a normalized sentence error rate as it doesn't depend as much on the definition of a word. As the metric which approximates user experience best we use WebScore: We send hypothesis and reference to a search backend and

compare the links we get back. Assuming that the reference generates the correct search result this way we know whether the search result for the hypothesis is within the first three results – such that the user can see the correct result on his smartphone screen.

Languages. After US English we launched Voice Search for the UK, Australia and India. Late 2009 Mandarin Chinese [5] and Japanese were added. Foreign languages pose many additional challenges. For example, some Asian languages like Japanese and Chinese don't have spaces between words. For these we wrote a segmenter which optimizes the word definitions maximizing sentence likelihood. Most languages have characters outside the normal ASCII set, in some cases thousands, which complicate automatic pronunciation rules.

Additional Challenges. There are many details which are critical to get right for a good user experience but we cannot discuss here because of space constraints. These include getting the user interface right, optimizing protocols for minimum latency, dealing with special cases like numbers, dates and abbreviations correctly, avoid showing offensive queries and improving the system efficiently after launch using the data coming in.

4 Outlook

For mobile devices speech is an attractive input modality and besides Voice Search we have been working on other features, including moer general Voice Input [6], contact dailing (as launched in the US) and recognition of special phrases to trigger certain applications on the phone. We believe that in the next few years speech input will become more accurate, more accepted and useful enough to help users efficiently access and navigate through information provided through mobile devices.

References

1. Bacchiani, M., Beaufays, F., Schalkwyk, J., Schuster, M., Strope, B.: Deploying GOOG-411: Early lessons in data, measurement, and testing. In: Proceedings of ICASSP, pp. 5260–5263 (2008)
2. van Heerden, C., Schalkwyk, J., Strope, B.: Language Modeling for What-with-Where on GOOG-411. In: Proceedings of Interspeech, pp. 991–994 (2009)
3. Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M., Strope, B.: Google Search by Voice: A case study. In: Weinstein, A. (ed.) Visions of Speech: Exploring New Voice Apps in Mobile Environments, Call Centers and Clinics. Springer, Heidelberg (2010) (in Press)
4. Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P., LeBeau, M.: Building transcribed speech corpora quickly and cheaply for many languages. In: Interspeech (submitted 2010)
5. Shan, J., Wu, G., Hu, Z., Tang, X., Jansche, M., Moreno, P.: Search by Voice in Mandarin Chinese. In: Interspeech (submitted 2010)
6. Ballinger, B., Allauzen, C., Gruenstein, A., Schalkwyk, J.: On-Demand Language Model Interpolation for Mobile Speech Input. In: Interspeech (submitted 2010)