Speak4it and the Multimodal Semantic Interpretation System

Michael Johnston [1], *Patrick Ehlen*[2]

[1] AT&T Labs Research, Florham Park, NJ, USA
[2] AT&T Labs, San Francisco, CA, USA
johnston@research.att.com, patrick.ehlen@att.com

Abstract

Multimodal interaction allows users to specify commands using combinations of inputs from multiple different modalities. For example, in a local search application, a user might say "gas stations" while simultaneously tracing a route on a touchscreen display. In this demonstration, we describe the extension of our cloud-based speech recognition architecture to a Multimodal Semantic Interpretation System (MSIS) that supports processing of multimodal inputs streamed over HTTP. We illustrate the capabilities of the framework using Speak4it[SM], a deployed mobile local search application supporting combined speech and gesture input. We provide interactive demonstrations of Speak4it on the iPhone and iPad and explain the challenges of supporting true multimodal interaction in a deployed mobile service.
Index Terms: multimodal, speech, gesture

1. Introduction

Commercial speech recognition and natural language processing systems have become commonplace over the past decade, and can now be accessed in places where speech recognition was not previously available. Likewise, the capabilities of gesture recognition systems have recently become more compact and inexpensive, and are now used in people's personal devices and homes. Combining these two technologies may help to create a new era of computer interaction in which people will at last be able to interact with computers using many of the same behaviors and methods they use to interact with each other, using language, gesture, and other natural methods of signaling [1].

This prospect of true multimodal interaction has motivated our expansion of our speech recognition platform [2] to support processing, recognition, and interpretation, not just of speech, but of multimodal input streams. The resulting Multimodal Semantic Interpretation System (MSIS) is designed to handle and process both low-level and high-level speech and gesture input signals, and to output semantic interpretations of these signals separately or in combination. Here we describe that system, and illustrate its use in Speak4it, the first commercially deployed multimodal application using the system.

2. Multimodal Semantic Interpretation System

MSIS was developed using AT&T's Watson platform, fostered in part by two advances: a streaming HTTP API that supports cloud-based recognition, and a plug-in development environment that supports complex chains of diverse processing for signal streams. While cloud-based speech recognition of streamed audio has become commonplace, Watson's MSIS can handle complex HTTP data streams that comprise not only audio, but also gesture, vision, or context data, or any other type of data one might wish to process. These data can be multiplexed into a single threaded data stream sent from a client application, including standardized data signal protocols like InkML [3] or EMMA [4], alongside custom protocols. Thus, an HTTP data connection to the cloud service has become a versatile pipe through which many types of data can be streamed simultaneously (Figure 1).

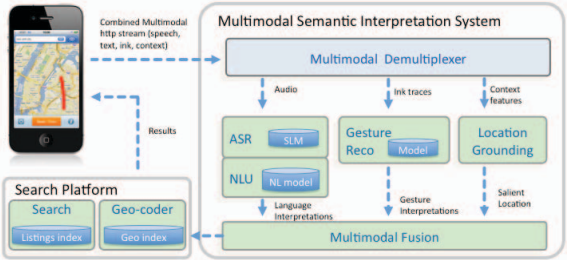

Figure 1. Multimodal architecture

At the other end of the HTTP connection, MSIS supports a chain of processing plug-ins that handles these streams of multiplexed data. The first stop for any stream is a de-multiplexing plug-in, which scans the HTTP input stream packets for different types of data and the boundaries that separate them. As these data arrive, they are passed on to buffers for handler plug-ins that begin the processing chain for the required type of data. Arrival of new data types triggers events that launch handlers for the new data. Thus de-multiplexed streams of data can undergo "sidechain" processing that handles one or more types of data, such as vision or gesture data, or a combination of the processing results from those data, such as composing lattices of semantic hypotheses of both ASR and gesture recognition.

To handle these different types of data, each Watson plug-in can load its own set of models, developed either using the tools provided with the recognizer, or using third-party tools (Watson supports plug-ins developed in C/C++ or Python). In addition to the acoustic and language models standard to ASR systems, there are now models for natural language understanding, gesture recognition, multimodal integration, context grounding, and disambiguation. Thus complex sets of combined streams of user interaction data can be processed in parallel and produce an integrated semantic interpretation of diverse and simultaneous signals. How does an application use this system to foster natural user interaction? To answer that question, we describe Speak4it, a freely available multimodal business search application for the iPhone and iPad [5].

3. Speak4it: Multimodal Local Search

Speak4it performs simultaneous recognition of user speech and deictic finger gestures performed on a digital map. In addition to providing this combined interpretation of speech and gesture, Speak4it also collects user context data gathered by the client device to help make disambiguation judgments about important but potentially ambiguous referents, like the user's intended location context. These data are streamed in

real-time along with the user's speech and finger gesture data, and analyzed by a context resolution plug-in in combination with the semantic output of both speech and gesture recognition.

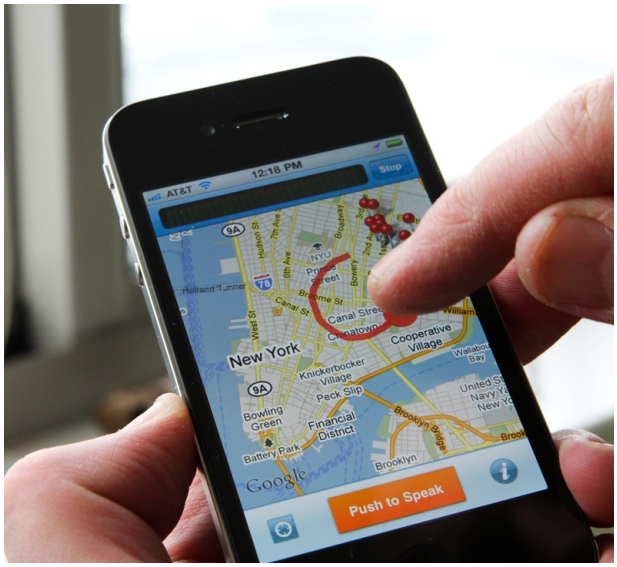

Figure 2. Speak4it[SM] interaction

On launching the application, users see a map of their area. They touch a "Push to Speak" button to initiate a query (Figure 2). They can speak any natural language combination of U.S. business name, business category, and U.S. geographic locations or landmarks, such as "Ninth Street Espresso" or "real estate attorneys in Detroit, Michigan" or "bicycle repair shops near the Golden Gate Bridge," and the map will zoom to the intended location and display the search results. Users can manipulate the map view using drag (pan) and pinch (zoom) touch gestures, or using spoken commands. If the user says "San Francisco," the map will zoom to display that city. They may also use multimodal commands where the location is specified directly by drawing on the map while speaking [6,7, 8], such as "Italian restaurants near here," in conjunction with a point or area gesture. A deictic point gesture returns results closest to the point, while an area gesture like a circle or polygon returns results within that area. A line gesture will produce results along that route.

As a user performs one of these multimodal commands, their speech audio, finger gesture movements, and device context data are multiplexed on the client and streamed over HTTP to the Watson MSIS, which de-multiplexes it and begins processing even as the user is still interacting with the device. The user's finger traces are passed to a gesture recognizer, which classifies the traces as a point, a line, or an area (in this case, an area). The audio stream is recognized using a statistical language model whose output is then passed to a natural language understanding plug-in (NLU) [9] that parses the query into semantic slots that include a topic phrase for the user's desired search subject (e.g., "Italian restaurants") and, if applicable, a location phrase (like "San Francisco") or a deictic location (like "here") that designates a desired location. In cases where there is an explicit location phrase (e.g. "pizza restaurants in San Francisco"), the location phrase is geo-coded so search results from the topic phrase may be sorted and displayed according to their proximity to the location. If the location is ambiguous, the location grounding plug-in receives context data streamed from the device and uses it in conjunction with semantic interpretation hypotheses from ASR and gesture to determine the user's intended search location [10]. At the end of the semantic processing chain, the intended search location is geocoded and sent with the search topic to a local search engine. The locations of relevant businesses are then displayed on the map. This local search system integrates simultaneous collection and processing of multiple streams of input data, and combines those behaviors from different types of user signals into a single semantic interpretation.

4. Conclusion

In the very near future, advanced user interfaces that recognize speech and gesture will require platforms that can receive and interpret multiple streams of information simultaneously, just as humans do when they communicate with each other face-to-face. When these platforms are designed as "interaction recognition" systems that handle different forms of data flexibly and close to the speech recognizer, they allow semantic interpretation to leverage speech lattices and basic recognition models to help inform other recognition tasks. We present a flexible and extensible multimodal semantic interpretation system that can deal with many forms of raw input signals and use them to produce integrated semantic interpretations. One application, Speak4it, already integrates simultaneous speech and gesture signal streams into a single semantic interpretation, and has been in use by the public for over a year. Future applications could extend beyond speech and finger gesture recognition to integrate streams of other types of kinesthetic and visual data.

5. Acknowledgements

Thanks to Jay Lieske, Clarke Retzer, Brant Vasilieff, Diamantino Caseiro, Junlan Feng, Srinivas Bangalore, Claude Noshpitz, Barbara Hollister, Remi Zajac, Mazin Gilbert, Barbara Hollister, and Linda Roberts for their contributions to Speak4it.

6. References

[1] Clark, H. H. and Krych, M. A., "Speaking While Monitoring Addressees for Understanding", J. of Memory and Language, 50: 62-81, 2004.

[2] Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tür, D., Ljolje, A., Parthasarathy, S., Rahim, M., Riccardi, G., and Saraclar, M., "The AT&T Watson Speech Recognizer", Proc. of ICASSP, 19-23, 2005.

[3] http://www.w3.org/2002/mmi/ink

[4] http://www.w3.org/TR/emma/

[5] http://speak4it.com/

[6] Cohen, P. R., Johnston, M., McGee, D., Oviatt, S.L., Pittman, J., Smith, I., Chen, L., and Clow, J., "Multimodal interaction for distributed interactive simulation", in M. Maybury and W. Wahlster [Eds], Readings in Intelligent Interfaces, 562-571, Morgan Kaufmann Publishers, 1998.

[7] Gustafson J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., and Wirén, M, "AdApt—A Multimodal Conversational Dialogue System in an Apartment Domain", Proc. of ICSLP, 2:134-137, 2000.

[8] Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., and Maloor, P, "MATCH: An Architecture for Multimodal Dialogue Systems. Proc. of ACL, 40:376-383, 2002.

[9] Feng, J., Bangalore, S., and Gilbert, M., "Role of Natural Language Understanding in Voice Local Search", Proc. of Interspeech, 1859-1862, 2009.

[10] Ehlen, P. and Johnston, M., Location Grounding in Multimodal Local Search. Proc. of ICMI-MLMI, 2010.