A Multimodal Dialogue Interface for Mobile Local Search

Patrick Ehlen
AT&T
525 Market Street
San Francisco, CA 94105
ehlen@research.att.com

Michael Johnston
AT&T Labs Research
180 Park Ave
Florham Park, NJ 07932
johnston@research.att.com

ABSTRACT
Speak4it[SM] uses a multimodal interface to perform mobile search for local businesses. Users combine simultaneous speech and touch to input queries or commands, for example, by saying, "gas stations," while tracing a route on a touchscreen. This demonstration will exhibit an extension of our multimodal semantic processing architecture from a one-shot query system to a multimodal dialogue system that tracks dialogue state over multiple turns and resolves prior context using unification-based context resolution. We illustrate the capabilities and limitations of this approach to multimodal interpretation, describing the challenges of supporting true multimodal interaction in a deployed mobile service, while offering an interactive demonstration on tablets and smartphones.

Author Keywords
Multimodal interfaces, Dialog, Search

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation (e.g. HCI): User Interfaces—input devices and strategies (e.g. mouse, touchscreen), natural language

General Terms
Algorithms, Design

INTRODUCTION
Speak4it is a mobile application for smartphones and tablets that provides a multimodal interface for users to find local business information. It offers true multimodal integration, allowing users to issue commands using simultaneous inputs from speech and touchscreen gestures traced with the finger on a dynamic map. For example, a user can say "gas stations" while tracing a route where they wish to search for gas stations, or, "Are there any nail salons in this area?" while circling a particular region.

When users interact with Speak4it, it streams speech, gesture, and context data to a server-based Multimodal Semantic Interpretation System (MSIS) [9]. There, speech and gesture recognition results are combined into a single semantic representation [1], and then interpreted in context to determine the query intended by the user.

Speak4it has been available as a multimodal search application [12] that performed traditional "one-shot" search, interpreting each new query independent of the context of prior queries. The only dialogue context it used was the user's last spoken location and map manipulations, which determined the intended search location from locations the user might have intended [5]. Disambiguation of the user's intention was informed by the theory of grounding [4], and leveraged some interaction context passed to the server with each new query. In this sense, Speak4it functioned as a rudimentary dialogue system.

But user query data collected over time showed users often expected the system to take more dialogue context into consideration, by making corrections or revisions to prior queries. Handling these dialogue moves calls for a more sophisticated approach to interaction, in the spirit of some other multimodal dialogue prototypes [cf. 7, 8, 10, 11].

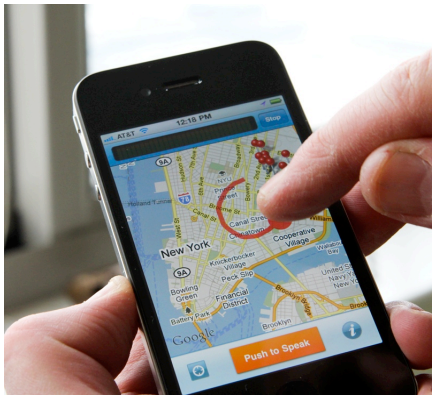

Figure 1. Speak4it[SM] interaction

MULTIMODAL DIALOGUE
The basic capabilities of Speak4it are to *Search*, such as "coffee shops around here"; and *Command*, such as "call this place". In this demonstration we illustrate and explain how Speak4it supports *Revisions*, e.g., "how about Korean?," and *Corrections*, "I meant California," using destructive unification.

For example, consider the following:

User: Japanese restaurants near Santa Monica?

[System returns some results near Santa Monica]

User: How about Chinese?

Copyright is held by the author/owner(s).
IUI'13 Companion, March 19–22, 2013, Santa Monica, CA, USA.
ACM 978-1-4503-1966-9/13/03

The full meaning of the refinement, "Chinese RESTAURANTS NEAR SANTA MONICA" results from combination of constraints from the dialog context with the refinement utterance, "How about Chinese?" To model this, we maintain an information state for each Speak4it client within the MSIS [9] platform. Part of this state encodes the last command represented as a typed feature structure:

$$\begin{bmatrix} term: & \begin{bmatrix} cuisine: & japanese \\ type: & restaurant \end{bmatrix} \\ location: & \begin{bmatrix} city: & santa\_monica \\ state: & california \\ type: & city\_state \end{bmatrix} \end{bmatrix}$$

When a refinement query, such as "How about Chinese?" is received, it is combined with the current context using a destructive unification operation similar to overlay [1] and extended with types. The destructive unification operation requires unifying feature structures to be of compatible types; and if they are not, then the newer feature structure replaces that of the previous turn. In this case, the refinement query is of a general type:

$$\begin{bmatrix} term: & \begin{bmatrix} cuisine: & chinese \\ type: & general \end{bmatrix} \end{bmatrix}$$

The unification result replaces the *cuisine* feature with "Chinese" and Speakit displays "Chinese restaurants near Santa Monica":

$$\begin{bmatrix} term: & \begin{bmatrix} cuisine: & chinese \\ type: & restaurant \end{bmatrix} \\ location: & \begin{bmatrix} city: & santa\_monica \\ state: & california \\ type: & city\_state \end{bmatrix} \end{bmatrix}$$

If instead the user said, "What about McDonalds," the resulting *term* feature is not type-compatible and the whole *term* value is replaced.

Similar operations model corrections, such as the following example, where the location is corrected:

User: chiropractors in Glendale

[System returns chiropractors in Glendale, California]

User: I meant Arizona

[System returns chiropractors in Glendale, Arizona]

Deictic gestures can also serve as location entities, so the user above might then say:

User: What about here? [circles an area]

[System zooms in to the circled region and returns chiropractors there]

In this multimodal correction example, the feature structure assigned to the multimodal combination of 'here' and the deictic gesture is assigned a type that is incompatible with *city_state*, and so the *location* feature of the refinement replaces the whole location in the structure.

REFERENCES

1. Alexandersson, J. and T. Becker. 2001. Overlay as the Basic Operation for Discourse Processing in a Multimodal Dialog System. Proceedings of IJCAI-01 Workshop on Knowledge and Reasoning in Practical Dialog Systems.
2. Bangalore, S. and Johnston, M. Robust understanding in multimodal interfaces. *Computational Linguistics, 35,* 3 (2009), 345-397.
3. Bocchieri, E., Caseiro, D., and Dimitriadis, D. Speech recognition modeling advances for mobile voice search. In *Proc. ICASSP 2011,* IEEE Press (2011), 4888-4891
4. Clark, H.H. *Using Language*. Cambridge University Press, Cambridge, MA, USA, 1996.
5. Ehlen, P. and Johnston, M. Location grounding in multimodal local search. In *Proc. ICMI-MLMI 2010*, ACM Press (2010).
6. Feng, J., Bangalore, S., and Gilbert, M. Role of natural language understanding in voice local search. In *Proc. Interspeech 2009*, ISCA (2009), 1859-1862.
7. Gustafson J., L. Bell, J. Beskow, J. Boye, R. Carlson, J. Edlund, B. Granström, D. House D., and M. Wirén. AdApt—A multimodal conversational dialogue system in an apartment domain. In *Proc. ICSLP 2000*, ISCA (2000), 134-137.
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. In *Proc. ACL 2002*, ACL (2002), 376-383.
9. Johnston, M. and P. Ehlen. Speak4it and the multimodal semantic interpretation system. In *Proc. Interspeech 2011*, ISCA (2011), 3333-3334.
10. Lemon, O., Bracy, A., Gruenstein, A., and Peters, S. 2001. Information states in a multi-modal dialogue system for human-robot conversation. In *Proc. Bi-Dialog 2001*, (2001), 57-67.
11. Wahlster, W. *SmartKom: Foundations of Multimodal Dialogue Systems (Cognitive Technologies)*. Springer-Verlag, New York, NY, USA, 2006.
12. http://speak4it.com/