Dilek Hakkani-Tür,
Gokhan Tur, and Larry Heck

Research Challenges and Opportunities in Mobile Applications

Following the rapid proliferation of mobile devices, especially smartphones, both server-based and embedded speech and multimodal applications started to emerge. These range from simpler applications where speech recognition is followed by a known task such as voice search or messaging, to more complex systems such as speech-to-speech translation, educational applications, or personal assistants. With the advances in client-side capabilities coming with larger screens that enable multitouch displays, these applications have begun to move beyond conventional speech applications towards three orthogonal dimensions: multimodality, personalization, and continuous and ubiquitous situation awareness. This has a significant effect, not only on the input side (for example, by allowing simultaneous touch and talk), but also on the output side, by making alternative types of information presentation (such as showing a map or Web page while displaying text or playing audio prompts) available.

The mobile device is no longer an island device but is rather fully connected to other personal devices [such as a personal computer (PC), tablet, or television (TV)] and data/information, enabling seamless and continuous transition between devices, where each device simply provides a different modality and view into a central repository of data/information in the cloud.

In this article, we highlight some of the technical challenges and research needed for mobile multimodal applications, especially focusing on educational aspects and research problems, pointing out issues and opportunities for students in this area. Our goal is to explore challenges, possibilities, and approaches for enabling speech processing, as well as convenient and effective speech and multimodal user interfaces for mobile environments.

Digital Object Identifier 10.1109/MSP.2011.941132
Date of publication: 15 June 2011

DATA-DRIVEN TECHNIQUES BEGAN TO DOMINATE THE FIELD FOLLOWING THE ADVANCES IN STATISTICAL MACHINE LEARNING METHODS.

FROM SPEECH TO MULTIMODAL

Figure 1 presents a very high-level timeline for speech-based and multimodal interactive systems. The earliest large scale systems, commonly known as interactive voice response (IVR) systems, are typically machine-directed dialog systems that ask users specific questions and expect user input to belong to a set of predetermined keywords or phrases [1].

The second generation spoken dialog systems allowed users to talk more naturally. Such systems were first developed under government-funded projects, such as the DARPA Communicator and ESPRIT Sundial projects [2], as well as within commercial research labs such as AT&T's How May I Help You (HMIHY)? system [3]. These systems typically include large vocabulary speech recognition, natural language understanding and generation, and dialog management components [4]. Data-driven techniques began to dominate the field following the advances in statistical machine learning methods.

MULTIMODAL CONVERSATIONAL INTERACTION FRAMEWORK

Although there were earlier multimodal interactive systems, such as AT&T MATCH, Microsoft MiPad, or ESPRIT MASK ([5]–[7], among others), the boom of smartphones resulted in a completely new array of multimodal applications. At a very high level, the basic components of such a system are shown in Figure 2, following [8]. This schema can also be adapted to interactive speech-to-speech translation, voice search, or dictation systems without loss of generality. Each interaction with the user is called a turn, and at each turn a user's input, $X_i = \{T_i, S_i, G_i\}$, can be in text (T_i), speech (S_i), and/or gestures (G_i). The goal of multimodal understanding is to then convert this multimodal input into a task-specific semantic representation of the user's intention A_u. This step involves recognition, semantic parsing, multimodal fusion, and interpretation. Interpretation can exploit semantic context C_u, such as the belief state of the system (which may include a user goal and actions, as well as dialog history), user specific meta-information, such as geolocation and personal preferences, and other contextual information. For example, if the user clicks on a map on the screen and asks "how much is the cheapest gas around here?" the system should be able to interpret the intent and the associated arguments, such as

```
Get_Price(good=gas,cost relative
  =cheapest, location=(latitude,
  longitude)).
```

More formally, statistical approaches estimate A_u as

1053-5888/11/$26.00©2011IEEE

$$\hat{A}_u = \underset{A_u}{\text{argmax}} \; P(A_u | X_i, C_u)$$
$$= \underset{A_u}{\text{argmax}} \; P(A_u | T_i, S_i, G_i, C_u).$$

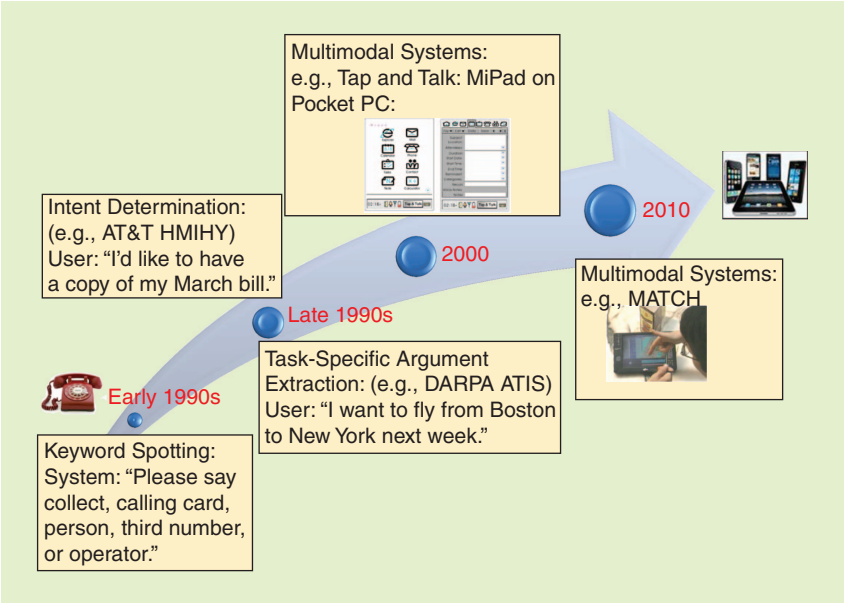
The dialog manager then decides on the most appropriate system action A_s. In the statistical approaches, this decision is made based on the expected reward over belief states, where the belief states are estimated using the previous machine action and belief state, as well as the observations the machine receives from the user [9]. However, all the traditional dialog management concepts need to be tailored to multimodal interaction from grounding to belief state optimization to response generation. For example the difference between implicit and explicit confirmations gets blurry with graphical user interface elements. Moreover, with the multimodal output, several different actions can be performed simultaneously. Hence, a response is generated similarly in a multimodal fashion, outputting $X_o = \{T_o, S_o, D_o\}$, that includes text (T_o), speech (S_o), and/or display elements (D_o), given the system's belief state concepts, C_s and the system action A_s.

$$\hat{X}_o = \underset{X_o}{\text{argmax}} \; P(X_o | A_s, C_s)$$
$$\{\hat{T}_o, \hat{S}_o, \hat{D}_o\} = \underset{T_o, S_o, D_o}{\text{argmax}} \; P(T_o, S_o, D_o | A_s, C_s).$$

For instance, if the dialog manager decides to present the information about the gas stations with prices, the multimodal generation needs to decide whether to show a map or list and/or speak this information to the user.

TOWARDS UBIQUITOUS AND CONTINUOUS PERSONAL INTERACTION

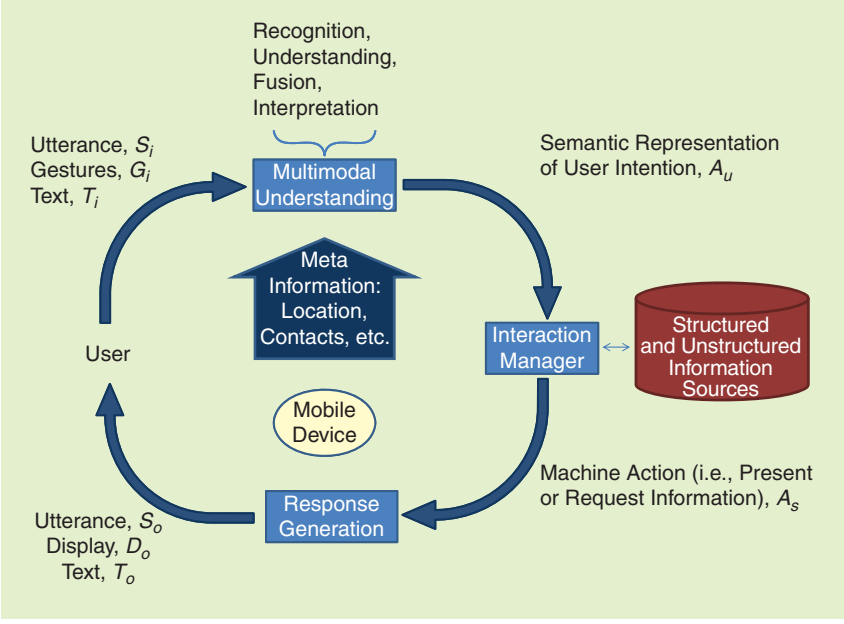
When telecommunications started moving from conventional land lines to smartphones allowing for graphical user interfaces, this was seen by some as the end of the speech-based applications, such as IVR. However, this has actually been the dawn of an exciting new era, where speech input is critical in enabling a number of technological opportunities associated with specific research challenges as described below.



[FIG1] **A brief history of multimodal interactive systems.**

PERSONALIZATION

Interactive systems are now becoming highly personalized. The mobile device can record vast amounts of private information about the user, ranging from contact lists and calendars to history of geolocations coupled with time and previous written or spoken communications. This user-specific metadata is waiting to be exploited for "customized" user interaction instead of "one size fits all" speech systems, such as voice search or call routing. Personalized mobile systems also enable longitudinal user studies, since users typically continue to use their mobile multimodal applications over a long period of time. One research challenge would be using



[FIG2] **A conceptual architecture of multimodal interactive systems.**

implicit supervision for user adaptation of models, exploiting user behavior patterns.

POWER OF THE CLOUD

The advances in cloud computing offer mobile application users an abstract view of multimodal processing services. All personal data, models, and application-specific information can be stored in the cloud. This allows users to interact with the system independent of the specific mobile device they have and enable developers to build and upgrade ubiquitous applications living in the cloud, alleviating the need for maintaining or updating servers or storage spaces. This comes at the cost of latency, which is often negligible to the users, provided streaming speech processing via the data channel.

POWER OF CONNECTIVITY TO OTHER DEVICES

The existence of the cloud enables the notion that the mobile phone is no longer an isolated device but rather fully connected to our other devices. This results in interesting research areas in building applications independent of mobile devices. Given that many personal devices (such as TVs, PCs, and tablets) or mobile devices are already connected to the Internet, there is no reason why the interaction is limited to only mobile. One can switch seamlessly from one screen to another and continue an interaction, since all the data maintenance and processing may be done on the cloud, provided client-side applications for the possible devices. While this is technically possible, the intelligent application must handle such screen transitions. Each device may simply provide a different experience and modality. The usage patterns and user behavior may vary over multiple screens, making the modeling of all components more challenging.

BIDIRECTIONAL DATA/VOICE

The setup of continuous communication channels gives an opportunity for a nearly "open-microphone" feel. In other words, instead of push-to-talk, one can simply interact with the system when needed, and similarly the system can proactively engage in a dialog with the user. Even when the user is not interacting with the system, the system can optionally be in "listening" mode, if needed. Such a setup enables many applications beyond interactive intelligent systems, such as health assessment or tracking systems. There are a number of significant challenges researchers need to tackle, ranging from understanding when the user is interacting with the system to robust recognition and processing of speech.

THE ADVANCES IN CLOUD COMPUTING OFFER MOBILE APPLICATION USERS AN ABSTRACT VIEW OF MULTIMODAL PROCESSING SERVICES.

RAPID PROTOTYPING AND BOOTSTRAPPING FOR DATA

Given that most speech applications require in-domain data, mobile applications provide an unmatched framework for this purpose. Currently, there are many sites that provide basic functionality, such as server-based speech recognition or synthesis. These capabilities make the engineering of a new mobile application easy and allow for a quick start to data collection for the specific domain/application/language of interest.

PROGRAMMABLE WEB

In most interactive systems, building the back-end infrastructure is a great engineering hurdle. The programmable Web offers solutions towards scalability for porting to new domains providing application programming interfaces (APIs) to knowledge sources and databases. The structure used in the back end may also drive the design of the semantic infrastructure and interaction flow. The biggest challenge is compensating for the misalignments between the available APIs and the functionalities of the system via natural interaction.

CONCLUSIONS

We have attempted to distill the research challenges and opportunities for mobile applications, ranging from personalization to connectivity. We believe that interactive multimodal mobile applications are still in their infancy and this is an exciting emerging field for research and development.

AUTHORS

Dilek Hakkani-Tür (dilek@ieee.org), ***Gokhan Tur*** (gokhan.tur@ieee.org), and ***Larry Heck*** (lheck@microsoft.com) are with Microsoft Speech Labs, Microsoft Research, in Mountain View, California.

REFERENCES

[1] J. G. Wilpon, L. R. Rabiner, C.-H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans. Speech Audio Processing*, vol. 38, no. 11, pp. 1870–1878, 1990.

[2] J. Peckham, "Speech understanding and dialogue over the telephone, an overview of the ESPRIT SUNDIAL project," in *Proc. DARPA Workshop on Speech and Natural Language*, Pacific Grove, CA, Feb. 1991.

[3] A. L. Gorin, A. Abella, T. Alonso, G. Riccardi, and J. H. Wright, "Automated natural spoken dialog," *IEEE Computer Mag.*, vol. 35, no. 4, pp. 51–56, Apr. 2002.

[4] G. Tur and R. De Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Hoboken, NJ: Wiley, 2011.

[5] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "MATCH: An architecture for multimodal dialogue systems," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July 2002.

[6] X. D. Huang, A. Acero, C. Chelba, L. Deng, J. Droppo, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, R. Loynd, M. Mahajan, P. Mau, S. Meredith, S. Mughal, S. Neto, M. Plumpe, K. Steury, G. Venolia, K. Wang, and Y. Wang, "Mipad: A multimodal interaction prototype," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT, May 2001.

[7] L. Lamel, S. Bennacef, J. Gauvain, H. Dartigues, and J. Temem, "User evaluation of the mask kiosk," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998.

[8] S. Young, "Talking to machines (statistically speaking)," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Denver, CO, Sept. 2002.

[9] J. D. Williams, P. Poupart, and S. Young, "Partially observable Markov decision processes with continuous observations for dialogue management," in *Proc SIGDIAL Workshop on Discourse and Dialogue*, Lisbon, Portugal, 2005.

[SP]