INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 8, 127–132, 2005

Multimodal Interfaces for Cell Phones and Mobile Technology

SHARON OVIATT AND REBECCA LUNSFORD
Center for Human-Computer Communication, Department of Computer Science & Engineering, Oregon Health & Science University, 20000 NW Walker Road, Beaverton, OR 97006, USA
oviatt@cse.ogi.edu
rebeccal@cse.ogi.edu

Abstract. By modeling users' natural spoken and multimodal communication patterns, more powerful and highly reliable interfaces can be designed that support emerging mobile technology. In this paper, we highlight three different examples of research that is advancing state-of-the-art mobile technology. The first is the development of fusion-based multimodal systems, such as ones that combine speech and pen or touch input, which are substantially improving the robustness and stability of system recognition. The second is modeling of multimodal communication patterns to establish open-microphone engagement techniques that work in challenging multi-person mobile settings. The third is new approaches to adaptive processing, which are able to transparently guide user input to match system processing capabilities. All three research directions are contributing to the design of more reliable, usable, and commercially promising mobile systems of the future.

Keywords: multimodal fusion, mobile interfaces, open microphone engagement, cognitive modeling, adaptive speech processing

Multimodal systems process two or more combined user input modes—such as speech, pen, touch, manual gestures, or gaze—in a coordinated manner with multimedia system output. The growing interest in multimodal interface design is inspired largely by the goal of supporting more transparent, flexible, efficient, robust, and powerfully expressive means of human-computer interaction (Oviatt, 2003). Such interfaces can be easier to learn and use, and are preferred for many applications because they give users greater control and flexibility in different situations. They also are preferred when tasks become difficult, and have been associated with a reduction of cognitive load and human performance errors (Oviatt et al., 2004). Multimodal systems have the potential to expand computing to more challenging applications, to be used by a broader spectrum of everyday people, and to accommodate more adverse usage conditions including many mobile contexts (Oulasvirta et al., 2005; Oviatt, 2000). As will be discussed in this paper, since a primary challenge in developing new recognition-based systems continues to be error handling, one key advantage of multimodal systems is simply their greater robustness compared with unimodal speech systems (Oviatt, 2002).

In this paper, we highlight three different examples of research that is advancing state-of-the-art mobile systems. The first is the development of fusion-based multimodal systems, such as ones that combine speech and pen input, which are substantially improving both the robustness and stability of recognition-based systems. The second is modeling of multimodal communication patterns to establish open-microphone engagement techniques that work in challenging mobile settings. The third is new approaches to adaptive processing, which are able to transparently guide user input to match system processing capabilities. All three of these directions are contributing to the design of more reliable, usable, and commercially promising mobile systems of the future. One general theme of this paper is that modeling users' natural spoken and multimodal communication patterns can yield more

powerful and reliable interfaces in support of emerging mobile systems.

Taming Mobile Recognition Errors: Multimodal Fusion-Based Systems

The performance advantage of a multimodal system that fuses two or more information sources can be substantial, for example resulting in excess of 40% robustness improvements. Such advantages have been documented for different modality combinations, varied tasks, and different environmental settings (Oviatt, 2002; Potamianos et al., 2004). Basically, recent research has shown that a well designed multimodal system can support *mutual disambiguation* of two input signals, which can produce a higher average likelihood of correct recognition and also more stable performance than one error-prone recognition technology alone (Oviatt, 1999, 2002). Figure 1 illustrates an example of mutual disambiguation from a system log in which a mobile user said, "zoom out" and drew a checkmark (Oviatt, 2000). Although the correct speech choice only was ranked fourth on the speech n-best list, nonetheless this semantic interpretation was recovered successfully on the final multimodal list. This occurred because inappropriate lexical hypotheses were weeded out when the system attempted to unify semantically-incompatible partial information from each input source.

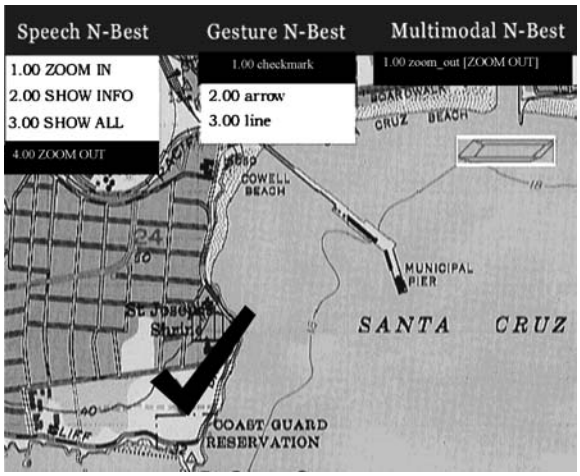

Figure 1. Incorrect speech interpretation is pulled up from fourth place on *n*-best list to achieve correct final multimodel interpretation in a fusion-based system due to mutual disambiguation.

In realistic noisy environments, such as work settings and in cars, speech-only recognition rates typically degrade substantially (e.g., 30–50%). This continues to pose the primary obstacle to widespread commercialization of spoken language technology. One study compared the performance of a multimodal system (i.e., speech and pen input) with a unimodal one (i.e., speech-only input) while users were in a quiet stationary setting versus walking around in a naturalistic noisy one (Oviatt, 2000). This study also evaluated the full range of microphone technology, from a high-end microphone (i.e., close-talking, directional, noise-canceling), to a low-end one (i.e., built-in on hand-held computer). As expected, the speech recognition error rate increased while users were mobile. These failures were attributed not only to environmental noise, but also to the adverse impact of users' reflexive Lombard adaptations while they were using the system mobile in the noisy setting.

The results of this mobile study revealed that multimodal system processing significantly improved system reliability over unimodal speech processing, both when users were stationary and when they were mobile. However, the largest improvements (i.e., approximately double) occurred when users were mobile. Compared with speech recognition, multimodal processing of speech and pen input decreased recognition failures overall by 19–35% when using the high- versus low-end microphones, respectively (Oviatt, 2000). These findings indicate why promising but error-prone recognition technologies such as speech are increasingly likely to be embedded within multimodal systems in the future. The net impact is a higher average level of system robustness, and also greater system stability as mobile users change settings between quiet and noisy ones. This kind of substantial reliability improvement will be essential for the commercial success of mobile systems that incorporate advanced recognition-based capabilities.

From the growing literature on error suppression, a few general design strategies can be distilled for optimizing multimodal system robustness (for others, see Oviatt, 2002):

- Increase the number of input modes or information sources incorporated within the multimodal system (e.g., trimodal systems perform more robustly than bimodal ones, which outperform unimodal ones)
- Combine input modes that represent semantically-rich information sources (e.g., incorporate pen-based

recognition of digits, rather than just selection or pointing)
- Increase the heterogeneity of information sources in the multimodal interface (e.g., fuse both behavioral and physiological information sources to identify a user, rather than just physiological ones)

One theme that emerges in these design guidelines is that whenever information is too scant or ambiguous for the system to recognize accurately (e.g., monosyllabic input, accented speaker, noisy mobile context), a multimodal system potentially can fortify robustness (Oviatt, 2002).

Are you Talking to me? Open-Microphone Engagement for Mobility

There currently is considerable interest in developing new microphone engagement techniques for advanced recognition systems that reside on mobile interfaces (e.g., cell phones, in-vehicle), which could leave users' eyes, hands, and attention freer for their primary task. In typical mobile tasks like name dialing on cell phones, interactions are brief and a considerable percentage of the interactive steps (e.g., 30–40%) simply involve engaging and disengaging the system. Furthermore, microphone engagement techniques currently require explicit user control, such as one or more key presses, a key press plus spoken keyword engagement, or pen-based "tap-to-talk." These techniques require the user to focus attention and cognitive resources on the mechanism of engagement per se, rather than on their primary field task. They also disrupt the otherwise "hands and eyes free" nature of many recognition-based applications. Such requirements can at times be unacceptably costly in terms of performance speed, accuracy, and a user's physical safety, especially in fluctuating real-world mobile settings. One aim of recent research has been to develop more *implicit* open-microphone engagement methods based on audio-visual processing of users' natural behavior, which have the potential to perform reliably without user distraction during difficult tasks, mobile use, and in other advantageous situations.

Newer audio-visual microphone engagement techniques typically include processing a user's head position as an estimate of gaze directed at a computer, along with the presence of articulated speech and corresponding lip movements. Using these rich processing techniques, the rate of speech/silence classification can be significantly improved over audio-only detection (Neti et al., 2000). However, recent research on human-robot interaction in a multi-person field setting indicates that head position and gaze actually are not reliable cues that a user is addressing the system rather than a human interlocutor (Katzenmaier et al., 2004). Other recent work on human-computer interaction likewise has shown that information on head position and gaze can completely fail to discriminate when a user is speaking to the system (Lunsford et al., in submission). On this topic, current audio-visual approaches remain rudimentary, and will need to be supplemented with additional information sources and more sophisticated user modeling.

In another line of work, two separate studies were conducted with adults 18–89 years of age. The results revealed that 65–80% of users produced self talk while interacting with the system, which was not directed to the system at all (Xiao et al., 2003; Lunsford et al., in submission). This highlights the fact that "noise" sources include not only ambient environmental noise and cross-talk among other users, but also a user's own self-talk and extraneous speech. In fact, in these studies people engaged in self talk before addressing the system over 30% of the time, with no decrease observed in younger adults compared with older ones (Lunsford et al., in submission). Self talk also increased steadily as users' task became more difficult. In these studies users looked at the system while engaging in self talk in virtually all cases, so once again head position and gaze would not have differentiated their self- from system-directed speech. As shown in Fig. 2, the results instead indicated that users' amplitude was a far more discriminating indicator of when their intended interlocutor was the computer. In fact, users' amplitude averaged a substantial 26 dBr[1] lower during self talk, and 96% of their self talk episodes could be discriminated from adjacent system-directed speech just using amplitude cues (Lunsford et al., in submission).

Achieving more reliable solutions for open-microphone engagement in real-world mobile settings clearly will require accurately distinguishing whether a user is talking to herself, one or more other persons, or a computer system. This ultimately will require more data collection and modeling of user-system interactions in different mobile usage contexts.

What Users Hear is What They Say: Adaptive Audio Interface Design for Mobility

The design of robust interfaces that process conversational speech is a challenging research direction in

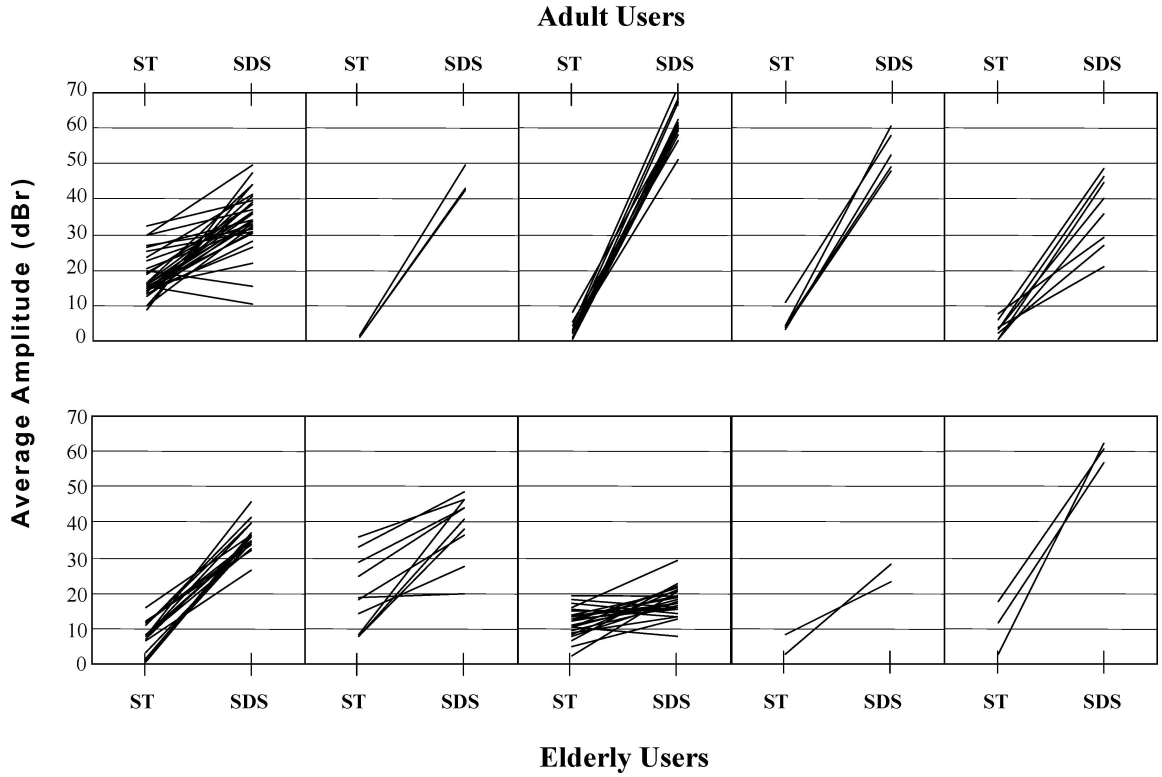

Figure 2. Individual differences for younger and elder adults in average amplitude for self talk (ST) versus system-directed speech (SDS) as matched utterance pairs.

large part because users' spoken language is so variable. However, human-computer interface research has revealed that the phenomenon of *linguistic convergence*, or the tendency of people's speech and language patterns to become more similar to those of their interactive partner, can be used to guide human input during system interactions. Within a text-based interface, Zoltan-Ford (1991) originally demonstrated that users' wordiness, lexical choice, and grammatical structure all are influenced directly by a system's prompts. In spoken and multimodal interfaces, text-to-speech (TTS) output also has been discovered to influence the acoustic-prosodic features of users' speech when they interact with a system (Oviatt et al., 2004). One implication of this work is that interface design can transparently guide user input to converge with different features of system output, thereby avoiding the use of explicit constraints on user behavior such as instruction, training, or error messages that may interrupt user-system interaction.

In a recent study involving twenty-four 7-to-10-year-old children, illustrated in Fig. 3, they conversed with

Figure 3. Eight-year old boy at school as he asks an animated marine character questions about itself.

animated characters that embodied different text-to-speech voices while learning about marine biology (Oviatt et al., 2004). An analysis of children's amplitude, duration, pause structure, dialogue response

Table 1. Magnitude change in different acoustic-prosodic features of users' speech as they converged on the computer text to-speech voice they heard.

Acoustic-prosodic feature	Magnitude change
Pause Duration	+49.0%
Number of Pauses	+26.4%
Amplitude	−22.4%
Dialogue Response Latency	+18.4%
Speech & Utterance Duration	+9.1–9.4%
Speech & Utterance Rate	+3.7–5.5%

[†]All features represent statistically significant change.

latencies, and other acoustic-prosodic features confirmed that they spontaneously adapted their speech 10–50% to more closely match the system's TTS output. As shown in Table 1, the largest adaptations involved pause structure and amplitude. These speech adaptations were rapid, and they also were dynamically readaptable whenever children were presented with a different computer character using a new TTS voice. In addition, these basic findings generalized across different subgroups of child users and different TTS voices. Although individual differences were evident in magnitude of adaptation, almost all users engaged in this speech adaptation.

In future mobile interfaces that must rely more on audio interface design (i.e., spoken and multimodal input, TTS and non-speech audio output), it is clear that this spontaneous convergence by users could be exploited to better guide their speech within system processing bounds, thereby enhancing robustness. Since current recognition technology remains sensitive to variations in amplitude and duration, future interfaces with TTS output could be designed to more actively manage these aspects of user input.

Although incorporation of TTS into interfaces has been viewed more as an art than a science, in the same research described above the auditory embodiment of animated characters as TTS output also had a significant impact on children's engagement in asking science questions. In particular, children asked 16% more science questions when conversing with characters that used a TTS voice resembling the speech of a master teacher (e.g., higher volume and pitch, wider pitch range), rather than other alternatives (Darves et al., 2004). These findings underscore the power of developing an interface with a task-appropriate metaphor, in this case instantiated exclusively by auditory means. More specifically, it reveals that conversational interfaces can be designed that effectively stimulate children during learning activities, thereby supporting the goals of next-generation educational software. As discussed in other recent work on auditory interface design (Cohen et al., 2004), the auditory personification of animated characters as social metaphors will be an especially important aspect of future mobility, and matching an appropriate TTS voice to an application domain can be an important tool for influencing user behavior.

Conclusion

New research directions were summarized on fusion-based multimodal systems, open-microphone engagement for field settings, and adaptive audio interface designs appropriate for mobility. This paper also highlighted the fact that empirically-based modeling of users' natural spoken and multimodal communication patterns has been an effective avenue for developing new mobile interface prototypes. Finally, interest in multimodal interface design has been inspired largely by the desire for more flexible, robust, and powerfully expressive means of human-computer interaction, as well as interface designs that are compatible with the growing demands of mobility.

Acknowledgments

This research was supported by DARPA Contract No. NBCHD030010 and NSF Grant No. IIS-0117868. Thanks to members of the Center for Human Computer Communication for assistance with conducting this research, as well as many helpful discussions. Table 1 and Fig. 1 reprinted with permission from ACM.

Note

1. 0 dBr defined as the ambient amplitude.

References

Cohen, M.H., Giangola, J.P., and Balogh, J. (2004). *Voice User Interface Design*. San Franciso, Ca: Addison-Wesley.

Darves, C. and Oviatt, S. (2004). Talking to digital fish: Designing effective conversational interfaces for educational software. In Z. Ruttkay and C. Pelachaud (Eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents*. Dordrecht: Kluwer, pp. 271–292.

Katzenmaier, M., Steifelhagen, R., and Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. ICMI*, pp. 144–151.

Lunsford, R., Oviatt, S., and Coulston, R. (Submission). Audio-visual cues distinguishing self- from system-directed speech in younger and older adults.

Neti, C., Iyengar, G., Potamianos, G., Senior, A., and Maison, B. (2000). Perceptual interfaces for information interaction: Joint processing of audio and visual information for human-computer interaction. In *Proc. ICSLP*, pp. 11–14.

Oulasvirta, A., Tamminen, S., Roto, V., and Kuorelahti, J. (2005). Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile HCI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'05), CHI Letters*. New York, N.Y.: ACM Press, to appear.

Oviatt, S.L. (1999). Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'99), CHI Letters*. New York, N.Y.: ACM Press, pp. 576–583.

Oviatt, S.L. (2000). Multimodal system processing in mobile environments. In *Proceedings of the Thirteenth Annual ACM Symposium on User Interface Software Technology (UIST'00), CHI Letters*. New York, N.Y.: ACM, pp. 21–30.

Oviatt, S.L. (2002). Breaking the robustness barrier: Recent progress on the design of robust multimodal systems. In M. Zelkowitz (ed.), *Advances in Computers*. Academic Press, vol. 56, pp. 305–341.

Oviatt, S.L. (2003). Multimodal interfaces. In J. Jacko and A. Sears (Eds.), *Handbook of Human-Computer Interaction*, Lawrence Erlbaum Assoc: Mahwah, New Jersey, chap. 14, pp. 286–304.

Oviatt, S.L., Coulston, R., and Lunsford, R. (2004). When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI'04)*, pp. 129–136.

Oviatt, S.L., Darves, C., and Coulston, R. (2004). Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *Transactions on Human Computer Interaction (TOCHI)*, *11*(3):300–328 (special issue on "Mobile and Adaptive Conversational Interfaces").

Potamianos, G., Neti, C., Luettin, J., and Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. In G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), *Issues in Visual and Audio-Visual Speech Processing*. Cambridge: MIT Press.

Xiao, B., Lunsford, R., Coulston, R., Wesson, R., and Oviatt, S.L. (2003). Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI'03)*. N.Y.: ACM Press, pp. 265–272.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, *34*:527–547.