More modes are better than one when it comes to comprehending human speech, especially when speakers are accented or interacting in noisy natural environments.

My focus here is recognition errors as a problem for spoken-language systems, especially when processing diverse speaker styles or speech produced in noisy field settings. However, when speech is combined with another input mode within a multimodal architecture, recent research has shown that two modes can function better than one alone. I also outline when and why multimodal systems display error-handling advantages. Recent studies on mobile speech and accented speakers have found that:

- Multimodal architectures combining speech and pen input can reduce speech recognition errors; and
- Improved robustness is greatest for the most challenging users and usage contexts.

In the future, multimodal systems will help stabilize error-prone recognition technologies, while also greatly expanding the accessibility of computing for everyday users and real-world environments.

Recognition Errors—The Achilles’ Heel of Speech Technology
Spoken-language systems involve recognition-based technology that is probabilistic in nature and therefore subject to misinterpretation. The Achilles’ heel limiting widespread commercializa-
tion of this technology is the rate of errors and lack of graceful error handling [8]. Benchmark error rates reported for speech recognition systems are still too high to support many applications [4]; the amount of time users spend resolving errors can be substantial and frustrating.

Although speech technology often performs adequately for native speakers of a language, for reading text, or when speaking in idealized laboratory conditions, current estimates indicate a 20%–50% decrease in recognition rates when speech is delivered under the following conditions:

- During natural spontaneous interaction;
- By diverse speakers (such as those with accents); or
- In a natural field environment.

Word-error rates are known to vary directly with speaking style, such that the more natural the speech delivery, the higher is the recognition system’s word-error rate [11]. In a study by Mitch Weintraub and his colleagues at SRI International, speakers’ word-error rates increased from 29% during carefully read dictation to 38% during a more conversationally read delivery, to 53% during natural spontaneous interactive speech. During spontaneous interaction, speakers are typically engaged in real tasks and generate variability in their speech for a number of reasons. For example, frequent miscommunication during a difficult task can prompt speakers to “hyperarticulate,” or speak in a more careful and clarified manner, leading to “durational” and other signal adaptations [8]. Interpersonal tasks or stress also can be associated with fluctuating emotional states, giving rise to pitch adaptations.

Basically, the recognition rate degrades whenever a user’s speech style departs in some way from the training data on which a recognizer was developed. Some speech adaptations, such as hyperarticulation, can be particularly difficult to process, because the signal changes begin and end abruptly and may affect only part of a longer utterance [8]. For handling speaker accents, a recognizer can be trained to recognize an individual accent, though it is far more difficult to successfully recognize varied accents (such as African, Asian, European, and North American) as might be required for an automated public telephone service or information kiosk. For handling heterogeneous accents, it can be infeasible to tailor an application to minimize highly confusable error patterns [6].

In addition to the difficulties presented by spontaneous speech and speakers’ stylistic adaptations, it is widely recognized that laboratory assessments overestimates recognition rates in natural field settings [2]. Field environments usually involve variable noise levels, social interchange, multitasking and interruption of tasks, increased cognitive load and human performance errors, and other sources of stress that have been estimated to produce 20%–50% drops in speech recognition accuracy. In fact, environmental noise is today viewed as a primary obstacle to the widespread commercialization of spoken-language technology [2, 3].

During field use and mobility, two main problems contribute to the degradation of system accuracy:

- Noise itself contaminates the speech signal, making it more difficult to process; and
- People speak differently in noisy conditions to make themselves understood.

“Stationary” noise sources (such as white noise) are often modeled and processed successfully, when they can be predicted, as in road noise in a moving car. However, many noises in natural field environments are “non-stationary” ones that either change abruptly or involve variable phase-in/phase-out noise as the speaker moves. Natural field environments also present qualitatively different sources of noise that cannot always be anticipated and modeled.

During noise, speakers also have an automatic normalization response, called the “Lombard effect,” that

Building a system that fuses two or more information sources can be an effective means of reducing recognition uncertainty, thereby improving robustness.
causes systematic speech modifications, including increased volume, reduced speaking rate, and changes in articulation and pitch [3]. The Lombard effect occurs not only in human adults, but in young children, primates, and even quail. From an interface-design perspective, it is important to note that the Lombard effect is essentially reflexive. As a result, it has not been possible to eliminate it through instruction or training or to suppress it selectively when noise is introduced [9].

Although speech originally produced in noise is actually more intelligible to a human listener, recognition accuracy degrades when a speech system processes Lombard speech, due to the increased departure between speech training and testing templates [3]. In addition to this difficulty handling Lombard speech, the template-matching approach used in current speech technology also has difficulty handling non-stationary sources of environmental noise.

Error Handling in Multimodal Interfaces
A different approach to resolving the impasse created by recognition errors is to design a more flexible multimodal interface incorporating speech as one of its input options. In the past, before robust multimodal approaches were available, skeptics believed that a multimodal system incorporating two error-prone recognition technologies (such as speech and handwriting recognition) would simply compound errors and yield even greater unreliability. However, recent data shows that building a system fusing two or more information sources can be an effective means of reducing recognition uncertainty, thereby improving robustness [5, 6, 10]. That is, the error-handling problems of recognition technologies typically become more manageable within a multimodal architecture.

User-centered error handling advantages. Previous research established that users prefer to interact multimodally and that their performance can be enhanced that way [7]. From a usability perspective, multimodal interfaces provide an opportunity for users to exercise their natural intelligence about when and how to use input modes effectively. First, in a multimodal interface, people avoid using an input mode they believe is error-prone for certain lexical content. For example, they are more likely to write than speak a foreign surname, relative to other content. Second, users’ language tends to be briefer and simpler linguistically when they interact multimodally, reducing the complexity of natural language processing and further minimizing errors [6]. Third, when a recognition error does occur, users alternate their choice of input modes in a manner that effectively resolves the error. Error resolution results because the confusion matrices typically differ for any particular lexical content for the two modes being recognized. These factors are all user-centered reasons why multimodal interfaces improve error avoidance and recovery, compared to unimodal spoken-language interfaces.

Architecture-based error-handling advantages. The increased robustness of multimodal systems also depends on an architecture that integrates modes synergistically. In a well-designed and optimized multimodal architecture, there can be “mutual disambiguation” of two input signals [6]. For example, Figure 1 shows mutual disambiguation from a user’s log during interaction with the QuickSet multimodal pen/voice system developed at the Oregon Graduate Institute. In this example, the user said “zoom out” and drew a checkmark. Although the lexical phrase “zoom out” was ranked fourth on the speech n-best list, the checkmark was recognized correctly by the gesture recognizer, and the correct semantic interpretation “zoom out” was recovered successfully and ranked first on the final multimodal n-best list. This recovery was achievable within the multimodal architecture because inappropriate signal pieces were discarded or “weeded out” during the unification process, which imposed semantic, temporal,
and other constraints on legal multimodal commands [7]. In this example, the three alternatives ranked higher on the speech n-best list integrate only with circle or question-mark gestures not present on the gesture n-best list. As a result, these alternatives could not form a legal integration and were weeded out.

It has been demonstrated empirically that by using the QuickSet architecture [1, 7, 12], a multimodal system can support mutual disambiguation of spoken and “gestural” input during semantic interpretation [5, 6]. As a result, such systems yield higher rates of correct interpretation than spoken-language processing alone. This performance improvement is the direct result of the disambiguation between signals that can occur in a well-designed multimodal system exhibiting greater stability and robustness.

Expanding Accessibility of Computing
One motivation for developing multimodal systems has been their potential for expanding the accessibility of computing to more diverse and non-specialist users while promoting new forms of computing not available in the past. There clearly are great differences in our individual abilities and desire to use different modes of communication, but multimodal interfaces are expected to increase the accessibility of computing for users of different ages, skill levels, and sensory and motor impairments. In part, a multimodal interface gives users interaction choices that can be used to circumvent personal limitations. A multimodal interface can be designed with a large performance advantage precisely for those users who tend to be the most disadvantaged by their reliance on speech input alone. An example is a non-native speaker whose accent is problematic for an English speech recognizer to process reliably.

Multimodal systems can also expand the usage contexts in which computing is viable to include, for example, natural field settings in which users are mobile. They permit users to alternate modes and switch between modalities as needed during the changing conditions of mobile use. Since speech and pen input are complementary along many dimensions, their combination provides broad utility across varied usage contexts. For example, a person may use hands-free speech input for voice-dialing a cell phone but switch to pen input to avoid speaking a business transaction in a noisy public place. A multimodal architecture can also be designed to enhance performance in less-than-ideal circumstances by adaptively weighting the alternative input modes in response to, for example, changing noise levels.

Multimodal performance for diverse accented speakers. In a recent study at the Oregon Graduate Institute, my colleagues and I evaluated the performance of the QuickSet pen/voice system to determine whether a multimodal architecture can be designed to support the following goals:

- Higher recognition rates than unimodal spoken-language processing:
Greater performance improvements for accented users;
Mutual disambiguation of incoming ambiguous signals; and
Better stability by using an alternate mode (pen) to disambiguate input that is unstable for certain user groups (speech).

The study’s participants were eight native speakers of English and eight accented speakers with native languages representing diverse continents, including Cantonese, Hindi, Mandarin, Spanish, Tamil, Turkish, and Yoruba. Everyone communicated 100 commands multimodally to the QuickSet system while using a handheld PC (see Figure 2). They set up simulation exercises involving community flood and fire management, as in the QuickSet interface in Figure 1. For example, they used speech and pen input to automatically locate objects and control the system’s map display, to add, orient, and move objects on the map, as well as to ask questions about the simulation and regulate system capabilities. Details of the QuickSet system’s functionality, interface design, signal and language processing, distributed agent-based framework, and symbolic/statistical hybrid architecture are outlined in [1, 7, 12].

A record of all user speech and pen input, along with the system’s performance, was recorded during 2,000 multimodal commands. This data was analyzed using the STAMP multimodal data logger and analysis tool, which calculates the system’s overall multimodal recognition rate [6]. This tool also identifies all recognition errors during speech and pen processing in each system module, including signal recognition, language interpretation, and multimodal integration phases. This evaluation includes all cases in which the correct lexical choice is not ranked first on the n-best list during any phase of system processing. It also includes all cases of mutual disambiguation in which a recognition failure occurs but the correct lexical choice is “retrieved” from lower down on its n-best list to produce a correct multimodal interpretation.

The study ultimately confirmed that the QuickSet multimodal architecture supports significant levels of mutual disambiguation, with one in eight user commands recognized correctly due to mutual disambiguation. Overall, the total error rate for spoken language was reduced 41% in the multimodal architecture, compared with spoken language processing as a standalone operation [6]. These results indicate that a multimodal system can be designed to function in a substantially more robust and stable manner than unimodal recognition-based technology.

Table 1. Accented and native speaker recognition-rate enhancement within a multimodal architecture.

<table>
<thead>
<tr>
<th>Type of Language Processing</th>
<th>% Performance Difference for Accented Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>−9.5% *</td>
</tr>
<tr>
<td>Gesture</td>
<td>+3.4% *</td>
</tr>
<tr>
<td>Multimodal</td>
<td>—</td>
</tr>
<tr>
<td>* significant difference</td>
<td></td>
</tr>
</tbody>
</table>

(b) Mutual disambiguation rate and ratio involving speech signal pull-ups for native and accented speakers.

<table>
<thead>
<tr>
<th>Type of Language Processing</th>
<th>Native Speakers</th>
<th>Accented Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutual disambiguation rate</td>
<td>8.5%</td>
<td>15.0% *</td>
</tr>
<tr>
<td>Ratio of speech pull-ups</td>
<td>.35</td>
<td>.65 *</td>
</tr>
</tbody>
</table>

* significant difference

Table 2. Mobile and stationary environment recognition-rate enhancement within a multimodal architecture.

(a) Difference in recognition rate performance in mobile environment, compared with stationary ones, for speech, gesture, and multimodal processing.

<table>
<thead>
<tr>
<th>Type of Language Processing</th>
<th>% Performance Difference when Mobile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>−10.0% *</td>
</tr>
<tr>
<td>Gesture</td>
<td>—</td>
</tr>
<tr>
<td>Multimodal</td>
<td>−8.0% *</td>
</tr>
<tr>
<td>* significant difference</td>
<td></td>
</tr>
</tbody>
</table>

(b) Mutual disambiguation rate and ratio involving speech signal pull-ups in stationary and mobile environments.

<table>
<thead>
<tr>
<th>Type of Language Processing</th>
<th>Stationary</th>
<th>Mobile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutual disambiguation rate</td>
<td>9.5%</td>
<td>16.0% *</td>
</tr>
<tr>
<td>Ratio of speech pull-ups</td>
<td>.26</td>
<td>.34 *</td>
</tr>
</tbody>
</table>

* significant difference

Table 1a confirms (as expected) that the speech-recognition rate was much poorer for accented speakers (−9.5%), though their gesture-recognition rate averaged slightly but significantly better (+3.4%). Table 1b shows that the rate of mutual disambiguation was significantly higher for accented speakers (+15%), compared with native speakers of English (+8.5%)—by a substantial 76%. As a result, Table
Multimodal performance in a mobile environment. In another study, my colleagues and I again investigated whether a multimodal architecture could support mutual disambiguation of input signals, as well as higher overall multimodal recognition rates than in spoken-language processing. We evaluated whether relatively greater performance gains would be produced in a noisy mobile environment, compared with a less error-prone quiet one. We were interested in determining whether the improved multimodal performance obtained with accented speakers was specific to that population, or whether general multimodal processing advantages apply to other challenging user groups and usage contexts.

In this study, 22 native English speakers interacted multimodally using the QuickSet pen/voice system on a handheld PC. Once again, the participants completed 100 commands in a procedure similar to the first study. However, each study participant completed 50 commands while working alone in a quiet room averaging 42 decibels (“stationary” condition) and another 50 commands while walking through a moderately noisy public cafeteria ranging from 40–60 decibels (“mobile” condition) (see Figure 3). Testing also involved two microphones, including a close-talking noise-canceling one and a built-in one lacking noise-cancellation technology. The study collected more than 2,600 multimodal utterances and evaluated performance metrics comparable to those in the first study.

One in seven utterances recognized correctly by the multimodal system was the result of mutual disambiguation, even though one or both component recognizers failed to identify the user’s intended meaning. This phenomenon also replicated across divergent microphones. In fact, 19%–35% reductions in the total error rate (for noise-canceling versus built-in microphones, respectively) were observed when speech was processed within the multimodal architecture. As in the first study, this substantial improvement in robustness was a direct result of the disambiguation between signals that can occur in multimodal systems.

Table 2a confirms (as expected) that the speech-recognition rate was significantly degraded while the same users were mobile in a naturalistic noisy setting (−10%), compared with stationary interaction in a quiet room. However, users’ gesture recognition rates did not decline during mobility, perhaps (partly) because pen-based input involves brief gestures, rather than extended handwriting. Table 2b shows that the mutual disambiguation rate also averaged substantially higher in the mobile condition (+16%), compared with stationary use (+9.5%). In fact, depending on which microphone was engaged, this rate ranged from 50%–100% higher during mobile system use. Table 2b also shows that failed speech signals likewise were “pulled up” more often by the multimodal architecture during mobile use.

Supporting robust recognition becomes extremely difficult when a spoken-language system has to process speech in naturalistic contexts involving variable sources and levels of noise, as well as qualitatively different types of noise (such as abrupt onset and phase-in/phase-out). Even when it is feasible to collect realistic mobile training data and to model many qualitatively different sources of noise, speech processing during abrupt shifts in noise (and the corresponding Lombard adaptations speakers make) is just plain difficult. As a result, mobile speech processing remains an unsolved problem for speech recognition. In the face of such challenges, a multimodal architecture supporting mutual disambiguation can potentially provide greater stability and a more viable long-term alternative for managing system errors.

Implications of these research findings. In both of these studies, even though one or both component recognizers failed to identify users’ intended meanings, the architectural constraints imposed by the QuickSet multimodal system’s semantic unification process ruled out incompatible speech and gesture integrations. These unification constraints effectively pruned recognition errors from the n-best lists of the component input modes, helping to significantly reduce the speech-recognition errors that are so prevalent for accented speakers and in mobile environments.

These studies together indicate that a multimodal architecture can stabilize error-prone recognition technologies (such as speech and pen input) to achieve improved robustness. As a result, next-generation multimodal systems can potentially harness new media—ones that are error-prone but expressively powerful and suitable for mobile use—within a more reliable architectural solution. While achieving this
important goal, they would also make technology available to a broader range of everyday users and usage contexts than has been possible before.

In both studies, speech recognition was the more unstable input mode, with most of the signal pull-ups retrieving poorly ranked speech input. There may be asymmetries in a multimodal interface as to which mode is the less reliable. In such cases, the most strategic approach for system development is to select an alternate mode that can act as a complement and stabilizer in promoting mutual disambiguation. In both studies, pen-based gestures fulfilled this purpose well, since their performance level was better for accented speakers and did not degrade in the mobile setting.

In multimodal research on speech and lip movements, similar robustness advantages have been documented for parallel processing of dual input signals [10]. For these modes, visually derived information about a speaker’s lip movements can improve recognition of the acoustic speech stream in a noisy environment. That is, spoken phonemes can be interpreted more reliably in the context of visible lip movements, or “visemes,” during noise. This type of multimodal system also provides a relatively greater boost in robustness as the noise level and phonemic speech-recognition errors increase [10].

More research is needed on how diverse users talk to speech-recognition systems during a wide variety of conditions—especially in realistic noisy environments where small devices are likely to be ubiquitous in the near future. Research is also needed on specific natural language, dialogue, adaptive processing, and other architectural techniques that optimize new multimodal systems further for mutual disambiguation and overall robustness.

Conclusion

The research outlined here demonstrates that multimodal architectures can stabilize error-prone recognition technologies, such as speech input, to yield substantial improvements in robustness. Although speech recognition as a standalone technology performs poorly for accented speakers and in mobile environments, the studies I cited showed that a multimodal architecture decreased failures in spoken-language processing by 19%–41%.

This performance improvement is due mainly to the mutual disambiguation of input signals that is possible within a multimodal architecture and that occurs at higher levels for challenging users and environments. These large robustness improvements can reduce or eliminate the performance gap for precisely those users and environments for whom and in which speech technology is most prone to failure. As a result, multimodal interfaces can be designed to expand the accessibility of computing—supporting diverse users in tangible ways and functioning more reliably during real-world conditions. During the next decade, we are increasingly likely to use expressive but error-prone new input modes embedded within multimodal architectures in a way that harnesses and stabilizes them more effectively.

References


Sharon Oviatt (oviatt@cse.ogi.edu) is a professor in the Department of Computer Science and Engineering at the Oregon Graduate Institute of Science and Technology and co-director of the Center for Human Computer Communication at OGI (www.cse.ogi.edu/CHCC).

This research was supported by Grant No. IRI-9350666 from the National Science Foundation, Special Exhibition for Creativity Grant No. IIS-9350666 from the National Science Foundation, Contracts DARPA-1-93-C-007 and N66001-99-D-8503 from DARPA’s Information Technology and Information Systems offices, Grant No. N00014-99-1-0577 from ONR, and by grants, gifts, and equipment donations from Boeing, Intel, Microsoft, and Motorola.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 2000 ACM 0002-0782/00/0900 $5.00