Breaking the Robustness Barrier: Recent Progress on the Design of Robust Multimodal Systems

SHARON OVIATT

Center for Human Computer Communication
Computer Science Department
Oregon Graduate Institute of Science and Technology
20,000 N. W. Walker Road
Beaverton, Oregon 97006
USA
oviatt@cse.ogi.edu

Abstract

Cumulative evidence now clarifies that a well-designed multimodal system that fuses two or more information sources can be an effective means of reducing recognition uncertainty. Performance advantages have been demonstrated for different modality combinations (speech and pen, speech and lip movements), for varied tasks (map-based simulation, speaker identification), and in different environments (noisy, quiet). Perhaps most importantly, the error suppression achievable with a multimodal system, compared with a unimodal spoken language one, can be in excess of 40%. Recent studies also have revealed that a multimodal system can perform in a more stable way than a unimodal one across varied real-world users (accented versus native speakers) and usage contexts (mobile versus stationary use). This chapter reviews these recent demonstrations of multimodal system robustness, distills general design strategies for optimizing robustness, and discusses future directions in the design of advanced multimodal systems. Finally, implications are discussed for the successful commercialization of promising but error-prone recognition-based technologies during the next decade.

ADVANCES IN COMPUTERS, VOL. 56
ISBN 0-12-012156-5

Copyright 2002 Elsevier Science Ltd
All rights of reproduction in any form reserved.

1. Introduction to Multimodal Systems

Multimodal systems process two or more combined user input modes—such as speech, pen, gaze, manual gestures, and body movements—in a coordinated manner with multimedia system output. This class of systems represents a new direction for computing, and a paradigm shift away from conventional windows–icons–menus–pointing device (WIMP) interfaces. Multimodal interfaces aim to recognize naturally occurring forms of human language and behavior, which incorporate at least one recognition-based technology (e.g., speech, pen, vision). The development of novel multimodal systems has been enabled by the myriad input and output technologies currently becoming available, including new devices and improvements in recognition-based technologies.

Multimodal interfaces have developed rapidly during the past decade, with steady progress toward building more general and robust systems [1,2]. Major developments have occurred in the hardware and software needed to support key component technologies incorporated within multimodal systems, and in techniques for integrating parallel input streams. The array of multimodal applications also has expanded rapidly, and currently ranges from map-based and virtual reality systems for simulation and training, to person identification/verification systems for security purposes, to medical and Web-based transaction systems that eventually will transform our daily lives [2–4]. In addition, multimodal systems have diversified to include new modality combinations, including speech and pen input, speech and lip movements, speech and manual gesturing, and gaze tracking and manual input [5–9].

This chapter specifically addresses the central performance issue of multimodal system design techniques for optimizing robustness. It reviews recent demonstrations of multimodal system robustness that surpass that of unimodal recognition systems, and also discusses future directions for optimizing robustness further through the design of advanced multimodal systems. Currently, there are two types of system that are relatively mature within the field of multimodal research, ones capable of processing users' speech and pen-based input, and others based

on speech and lip movements. Both types of system process two recognition-based input modes that are semantically rich, and have received focused research and development attention. As we will learn in later sections, the presence of two semantically rich input modes is an important prerequisite for suppression of recognition errors. The present chapter will focus on a discussion of these two types of multimodal system.

1.1 Types of Multimodal System

Since the appearance of Bolt's "Put That There" [10] demonstration system, which processed speech in parallel with touch-pad pointing, a variety of new multimodal systems have emerged. Most of the early multimodal systems processed simple mouse or touch-pad pointing along with speech input [11–16]. However, contemporary multimodal systems that process two parallel input streams, each of which is capable of conveying rich semantic information, have now been developed. These multimodal systems recognize two natural forms of human language and behavior, for which two recognition-based technologies are incorporated within a more powerful bimodal user interface.

To date, systems that combine either speech and pen input [2,17] or speech and lip movements [1,7,18] are the predominant examples of this new class of multimodal system. In both cases, the keyboard and mouse have been abandoned. For speech and pen systems, spoken language sometimes is processed along with complex pen-based gestural input involving hundreds of different symbolic interpretations beyond pointing [2]. For speech and lip movement systems, spoken language is processed along with corresponding human lip movement information during the natural audio–visual experience of spoken interaction. In both cases, considerable work has been directed toward quantitative modeling of the integration and synchronization characteristics of the two input modes being processed, and innovative new time-sensitive architectures have been developed to process these rich forms of patterned input in a robust manner. Recent reviews of the cognitive science underpinnings, natural language processing and integration techniques, and architectural features used in these two types of multimodal system have been summarized elsewhere (see Benoit *et al.* [1], Oviatt *et al.* [2], and Oviatt [19]).

Multimodal systems designed to recognize speech and pen-based gestures first were prototyped and studied in the early 1990s [20], with the original QuickSet system prototype built in 1994. The QuickSet system is an agent-based, collaborative multimodal system that runs on a hand-held PC [6]. As an example of a multimodal pen/voice command, a user might add three air landing strips to a map by saying "airplane landing strips facing this way (draws arrow NW), facing this way (draws arrow NE), and facing this way (draws arrow SE)." Other systems

of this type were built in the late 1990s, with examples including the Human-centric Word Processor, Portable Voice Assistant, QuickDoc, and MVIEWS [2,21–23]. In most cases, these multimodal systems jointly interpreted speech and pen input based on a frame-based method of information fusion and a late semantic fusion approach, although QuickSet uses a statistically ranked unification process and a hybrid symbolic/statistical architecture [24]. Other very recent speech and pen multimodal systems also have begun to adopt unification-based multimodal fusion and hybrid processing approaches [25,26], although some of these newer systems still are limited to pen-based pointing. In comparison with the multimodal speech and lip movement literature, research and system building on multimodal speech and pen systems has focused more heavily on diversification of applications and near-term commercialization potential.

In contrast, research on multimodal speech and lip movements has been driven largely by cognitive science interest in intersensory audio–visual perception, and the coordination of speech output with lip and facial movements [5,7,27–36]. Among the contributions of this literature has been a detailed classification of human lip movements (visemes), and the viseme–phoneme mappings that occur during articulated speech. Actual systems capable of processing combined speech and lip movements have been developed during the 1980s and 1990s, and include the classic work by Petajan [37], Brooke and Petajan [38], and others [39–43]. Additional examples of speech and lip movement systems and applications have been detailed elsewhere [1,7]. The quantitative modeling of synchronized phoneme/viseme patterns that has been central to this multimodal literature recently has been used to build animated characters that generate text-to-speech output with coordinated lip movements for new conversational interfaces [28,44]. In contrast with the multimodal speech and pen literature, which has adopted late integration and hybrid approaches to processing dual information, speech and lip movement systems sometimes have been based on an early feature-level fusion approach. Although very few existing multimodal interfaces currently include adaptive processing, researchers in this area have begun exploring adaptive techniques for improving system robustness during noise [45–47]. This is an important future research direction that will be discussed further in Section 2.2.2.

As multimodal interfaces gradually evolve toward supporting more advanced recognition of users' natural activities in context, including the meaningful incorporation of vision technologies, they will begin to support innovative directions in pervasive interface design. New multimodal interfaces also will expand beyond rudimentary bimodal systems to ones that incorporate three or more input modes, qualitatively different modes, and more sophisticated models of multimodal interaction. This trend already has been initiated within biometrics research, which has combined recognition of multiple behavioral input modes (e.g., speech,

handwriting, gesturing, and body movement) with physiological ones (e.g., retinal scans, fingerprints) in an effort to achieve reliable person identification and verification in challenging field conditions [4,48].

1.2 Motivation for Multimodal System Design

The growing interest in multimodal interface design is inspired largely by the goal of supporting more flexible, transparent, and powerfully expressive means of human–computer interaction. Users have a strong preference to interact multimodally in many applications, and their performance is enhanced by it [2]. Multimodal interfaces likewise have the potential to expand computing to more challenging applications, to a broader spectrum of everyday users, and to accommodate more adverse usage conditions such as mobility. As this chapter will detail, multimodal interfaces also can function in a more robust and stable manner than unimodal systems involving a single recognition-based technology (e.g., speech, pen, vision).

1.2.1 Universal Access and Mobility

A major motivation for developing more flexible multimodal interfaces has been their potential to expand the accessibility of computing to more diverse and nonspecialist users. There are large individual differences in people's ability and preference to use different modes of communication, and multimodal interfaces are expected to increase the accessibility of computing for users of different ages, skill levels, cultures, and sensory, motor, and intellectual impairments. In part, an inherently flexible multimodal interface provides people with interaction choices that can be used to circumvent personal limitations. This is becoming increasingly important, since U.S. legislation effective June 2001 now requires that computer interfaces demonstrate accessibility in order to meet federal procurement regulations [49,50]. Such interfaces also permit users to alternate input modes, which can prevent overuse and damage to any individual modality during extended computing tasks (R. Markinson, University of California at San Francisco Medical School, 1993).

Another increasingly important advantage of multimodal interfaces is that they can expand the usage contexts in which computing is viable, including natural field settings and during mobility. In particular, they permit users to switch between modes as needed during the changing conditions of mobile use. Since input modes can be complementary along many dimensions, their combination within a multimodal interface provides broader utility across varied and changing usage contexts. For example, a person with a multimodal pen/voice interface may use

hands-free speech input for voice dialing a car cell phone, but switch to pen input to avoid speaking a financial transaction in a public setting.

1.2.2 Error Avoidance and Resolution

Of special relevance to this chapter, multimodal interface design frequently manifests improved error handling, in terms of both error avoidance and graceful recovery from errors [43,51–55]. There are *user-* and *system-centered* reasons why multimodal systems facilitate error recovery, when compared with unimodal recognition-based interfaces. First, in a multimodal speech and pen interface, users will select the input mode that they judge less error prone for particular lexical content, which tends to lead to error avoidance [51]. For example, they may prefer speedy speech input, but will switch to pen input to communicate a foreign surname. Secondly, users' language often is simplified when interacting multimodally. In one study, a user added a boat dock to an interactive map by speaking

"Place a boat dock on the east, no, west end of Reward Lake."

When using multimodal pen/voice input, the same user completed the same action with

[draws rectangle] "Add dock."

Multimodal utterances generally were documented to be briefer, and to contain fewer disfluencies and complex locative descriptions, compared with a speech-only interface [56]. This can result in substantially reducing the complexity of natural language processing that is needed, thereby reducing recognition errors [57]. Thirdly, users have a strong tendency to switch modes after a system recognition error, which tends to prevent repeat errors and to facilitate error recovery. This error resolution occurs because the confusion matrices differ for any given lexical content for the two recognition technologies involved [52].

In addition to these user-centered reasons for better error avoidance and resolution, there also are system-centered reasons for superior error handling. A well-designed multimodal architecture with two semantically rich input modes can support *mutual disambiguation* of signals. For example, Fig. 1 illustrates mutual disambiguation from a user's log during an interaction with the QuickSet multimodal system. In this example, the user said "zoom out" and drew a checkmark. Although the lexical phrase "zoom out" only 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 (i.e., ranked first) on the final multimodal n-best list. As a result, the map interface

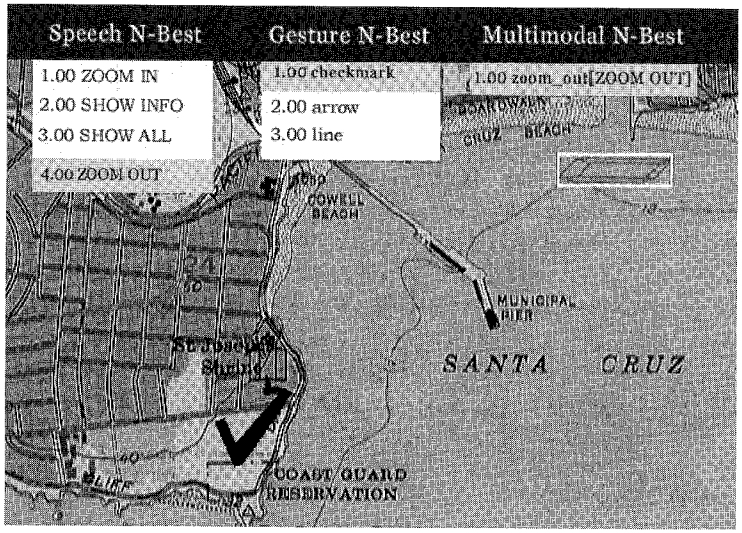

FIG. 1. QuickSet user interface during multimodal command to "zoom out," illustrating mutual disambiguation with the correct speech interpretation pulled up on its n-best list to produce a correct final multimodal interpretation.

zoomed out correctly, and no errors were ever experienced by the user. This recovery of the correct interpretation was achievable within the multimodal architecture because inappropriate signal pieces are discarded or "weeded out" during the unification process, which imposes semantic, temporal, and other constraints on what can be considered "legal" multimodal interpretations [2,6]. In this particular example, the three alternatives ranked higher on the speech n-best list only could have integrated with circle or question mark gestures, which were not present on the n-best gesture list. As a result, these alternatives could not form a legal integration and were discarded.

Using the QuickSet architecture, which involves late semantic integration and unification [2,6,24], it has been demonstrated empirically that a multimodal system can support mutual disambiguation of speech and pen input during semantic interpretation [53,58,59]. As a result, such a system yields a higher overall rate of correct utterance interpretations 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, because each mode provides context for interpreting the other during integration. To achieve optimal disambiguation of meaning, a multimodal interface ideally should be designed to include complementary input modes, and each mode should provide duplicate functionality such that users can accomplish their goals using either one.

Parallel error suppression also has been observed in multimodal speech and lip movement systems, although the primary focus has been on demonstrating improvements during noise. During the audio–visual perception of speech and lip movements, enhancement of multimodal speech recognition has been demonstrated over audio-only processing for human listeners [5,30,32,34,60] and also for multimodal speech and lip movement systems [3,39,43,45,61–65]. In this literature, key complementarities have been identified between acoustic speech and corresponding lip movements, which jointly supply unique information for accurately recognizing phonemes. More detailed research findings on the error suppression capabilities and mechanisms of multimodal systems will be reviewed in Section 2.2.

1.3 Long-Term Directions: Multimodal–Multisensor Systems That Model Biosensory Perception

The advent of multimodal interfaces based on recognition of human speech, gaze, gesture, and other natural behavior represents only the beginning of a progression toward computational interfaces capable of relatively human-like sensory perception. Such interfaces eventually will interpret continuous input from a large number of different visual, auditory, tactile, and other input modes, which will be recognized as users engage in everyday activities. The same system will track and incorporate information from multiple sensors on the user's interface and surrounding physical environment in order to support intelligent adaptation to the user, task and usage environment. This type of advanced multimodal–multisensor interface will be integrated within a flexible architecture in which information from different input modes or sensors can be *actively recruited* when it is relevant to the accurate interpretation of an ongoing user activity. The flexible collection of information essentially will permit *dynamic reconfiguration* of future multimodal–multisensor interfaces, especially when key information is incomplete or discordant, or at points when the user's activity changes.

Adaptive multimodal–multisensor interfaces that incorporate a broad range of information have the potential to achieve unparalleled robustness, and to support new functionality. They also have the potential to perform flexibly as multifunctional and personalized mobile interfaces. At their most evolved endpoint, this new class of interfaces will become capable of relatively human-like sensory-perceptual capabilities, including self-diagnostic functions.

The long-term research direction of designing robust multimodal–multisensor interfaces will be guided in part by biological, neurophysiological, and psychological evidence on the organization of intelligent sensory perception [66]. Coordinated sensory perception in humans and animals is active, purposeful, and able to achieve remarkable robustness through multimodality [5,30,32,34,67,68]. In

fact, robustness generally is achieved by integrating information from many different sources, whether different input modes, or different kinds of data from the same mode (e.g., brightness, color). During fusion of perceptual information, for example, the primary benefits include improved robustness, the extraction of qualitatively new perceptions (e.g., binocular stereo, depth perception), and compensation for perceptual disturbance (e.g., eye movement correction of perturbations induced by head movement).

In biological systems, input also is dynamically recruited from relevant sensory neurons in a way that is both sensitive to the organism's present context, and informed by prior experience [69–71]. When orienting to a new stimulus, the collection of input sampled by an organism can be reconfigured abruptly. Since numerous information sources are involved in natural sensory perception, discordant or potentially faulty information can be elegantly resolved by recalibration or temporary suppression of the "offending" sensor [72–74].

In designing future architectures for multimodal interfaces, important insights clearly can be gained from biological and cognitive principles of sensory integration, intersensory perception, and their adaptivity during purposeful activity. As a counterpoint, designing robust multimodal interfaces also requires a computational perspective that is informed by the implementation of past fusion-based systems. Historically, such systems often have involved conservative applications for which errors are considered costly and unacceptable, including biometrics, military, and aviation tasks [4,75,76]. However, fusion-based systems also have been common within the fields of robotics and speech recognition [3,7,18,24,39, 43,47,77]. Although discussion of these many disparate literatures is beyond the scope of this chapter, nonetheless examination of the past application of fusion techniques can provide valuable guidance for the design of future multimodal–multisensor interfaces. In the present chapter, discussion will focus on research involving multimodal systems that incorporate speech recognition.

2. Robustness Issues in the Design of Recognition-Based Systems

As described in the Introduction, state-of-the-art multimodal systems now are capable of processing two parallel input streams that each convey rich semantic information. The two predominant types of such a system both incorporate speech processing, with one focusing on multimodal speech and pen input [2,17], and the other multimodal speech and lip movements [1,7,18]. To better understand the comparative robustness issues associated with unimodal versus multimodal system design, Section 2.1 will summarize the primary error handling problems with unimodal recognition of an acoustic speech stream. Although spoken

language systems support a natural and powerfully expressive means of interaction with a computer, it is still the case that high error rates and fragile error handling pose the main interface design challenge that limit the commercial potential of this technology. For comparison, Section 2.2 will review research on the relative robustness of multimodal systems that incorporate speech. Section 2.3 then will summarize multimodal design strategies for optimizing robustness, and Section 2.4 will discuss the performance metrics used as forcing functions for achieving robustness.

2.1 Recognition Errors in Unimodal Speech Systems

Spoken language systems involve recognition-based technology that by nature is probabilistic and therefore subject to misinterpretation. Benchmark error rates reported for speech recognition systems still are too high to support many applications [78], and the time that users spend resolving errors can be substantial and frustrating. Although speech technology often performs adequately for read speech, for adult native speakers of a language, or when speaking under idealized laboratory conditions, current estimates indicate a 20–50% decrease in recognition rates when speech is delivered during natural spontaneous interactions, by a realistic range of diverse speakers (e.g., accented, child), or in natural field environments.

Word error rates (WERs) are well known to vary directly with speaking style, such that the more natural the speech delivery the higher the recognition system's WER. In a study by Weintraub *et al.* [79], speakers' WERs 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 typically are engaged in real tasks, and this generates variability in their speech for several reasons. For example, frequent miscommunication during a difficult task can prompt a speaker to hyperarticulate during their repair attempts, which leads to durational and other signal adaptations [80]. Interpersonal tasks or stress also can be associated with fluctuating emotional states, giving rise to pitch adaptations [81].

Basically, the recognition rate degrades whenever a user's speech style departs in some way from the training data upon which a recognizer was developed. Some speech adaptations, like hyperarticulation, can be particularly difficult to process because the signal changes often begin and end very abruptly, and they may only affect part of a longer utterance [80]. In the case of speaker accents, a recognizer can be trained to recognize an individual accent, although it is far more difficult to recognize varied accents successfully (e.g., Asian, European, African, North American), as might be required for an automated public telephone service. In the case of heterogeneous accents, it can be infeasible to specifically tailor an

application to minimize highly confusable error patterns in a way that would assist in supporting robust recognition [53].

The problem of supporting adequate recognition rates for diverse speaker groups is due partly to the need for corpus collection, language modeling, and tailored interface design with different user groups. For example, recent research has estimated that children's speech is subject to recognition error rates that are two-to-five times higher than adult speech [82–85]. The language development literature indicates that there are specific reasons why children's speech is harder to process than that of adults. Not only is it less mature, children's speech production is inherently more variable at any given stage, and it also is changing dynamically as they develop [86,87].

In addition to the many difficulties presented by spontaneous speech, speaker stylistic adaptations, and diverse speaker groups, it is widely recognized that laboratory assessments overestimate the recognition rates that can be supported in natural field settings [88–90]. Field environments typically involve variable noise levels, social interchange, multitasking and interruption of tasks, increased cognitive load and human performance errors, and other sources of stress, which collectively produce 20–50% drops in speech recognition accuracy. In fact, environmental noise currently is viewed as one of the primary obstacles to widespread commercialization of spoken language technology [89,91].

During field use and mobility, there actually are two main problems that contribute to degradation in system accuracy. The first is that noise itself contaminates the speech signal, making it harder to process. Stationary noise sources often can be modeled and processed successfully, when they can be predicted (e.g., road noise in a moving car). However, many noises in natural field environments are *nonstationary* ones that either change abruptly or involve variable phase-in/phase-out noise as the user moves. Natural field environments also present qualitatively different sources of noise that cannot always be anticipated and modeled. Speech technology has special difficulty handling abrupt onset and nonstationary sources of environmental noise.

The second key problem, which has been less well recognized and understood, is that people speak differently under noisy conditions in order to make themselves understood. During noise, speakers have an automatic normalization response called the "Lombard effect" [92], which causes systematic speech modifications that include increased volume, reduced speaking rate, and changes in articulation and pitch [58,91,93–95]. The Lombard effect not only occurs in human adults, but also in young children, primates, quail, and essentially all animals [96–98]. From an interface design standpoint, it is important to realize that the Lombard effect essentially is 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 [99].

Although speech originally produced in noise actually is *more* intelligible to a human listener, a system's recognition accuracy instead degrades when it must process Lombard speech [91].

To summarize, current estimates indicate a 20–50% decrease in recognition rate performance when attempts are made to process natural spontaneous speech, or speech produced by a wider range of diverse speakers in real-world field environments. Unfortunately, this is precisely the kind of realistic speech that must be recognized successfully before widespread commercialization can occur. During the development of modern speech technology there generally has been an overreliance on hidden Markov modeling, and a relatively singular focus on recognizing the phonetic features of acoustic speech. Until very recently, the speech community also has focused quite narrowly on unimodal speech processing. Finally, speech recognition research has depended very heavily on the word error rate as a forcing function for advancing its technology. Alternative perspectives on the successful development of robust speech technology will be discussed throughout this chapter.

2.2 Research on Suppression of Recognition Errors in Multimodal Systems

A different approach to resolving the impasse created by recognition errors is to design a more flexible multimodal interface that incorporates speech as one of the input options. In the past, skeptics have claimed that a multimodal system incorporating two error-prone recognition technologies would simply compound errors and yield even greater unreliability. However, as introduced earlier, cumulative data now clarify that a system which fuses two or more input modes can be an effective means of reducing recognition uncertainty, thereby improving robustness [39,43,53,58]. Furthermore, performance advantages have been demonstrated for different modality combinations (speech and pen, speech and lip movements), for varied tasks (map-based simulation, speaker identification), and in different environments (noisy mobile, quiet stationary). Perhaps most importantly, the error suppression achievable with a multimodal system, compared with an acoustic-only speech system, can be very substantial in noisy environments [39,45,58,62,64,65]. Even in environments not degraded by noise, the error suppression in multimodal systems can exceed 40%, compared with a traditional speech system [53].

Recent studies also have revealed that a multimodal architecture can support *mutual disambiguation* of input signals, which stabilizes the system's performance in a way that can minimize or even close the recognition rate gap between nonnative and native speakers [53], and between mobile and stationary system use [58]. These results indicate that a well-designed multimodal system not only can

perform overall more robustly than a unimodal system, but they also can perform in a more reliable way across varied real-world users and usage contexts. In the following sections, research findings that compare the robustness of multimodal speech processing with parallel unimodal speech processing will be summarized. Relevant studies will be reviewed on this topic from the multimodal literature on speech and pen systems and speech and lip movement systems.

2.2.1 Robustness of Multimodal Speech and Pen Systems

The literature on multimodal speech and pen systems recently has demonstrated error suppression ranging between 19 and 41% for speech processed within a multimodal architecture [53,58]. In two recent studies involving over 4600 multimodal commands, these robustness improvements also were documented to be greater for diverse user groups (e.g., accented versus native speakers) and challenging usage contexts (noisy mobile contexts versus quiet stationary use), as introduced above. That is, multimodal speech and pen systems typically show a larger performance advantage precisely for those users and usage contexts in which speech-only systems typically fail. Although recognition rates degrade sharply under the different kinds of conditions discussed in Section 2.1, nonetheless new multimodal pen/voice systems that improve robustness for many of these challenging forms of speech can be designed.

Research on multimodal speech and pen systems also has introduced the concept of mutual disambiguation (see Section 1.2 for definition and illustration). This literature has documented that a well-integrated multimodal system that incorporates two semantically rich input modes can support significant levels of mutual disambiguation between incoming signals. That is, a synergistic multimodal system can be designed in which each input mode disambiguates partial or ambiguous information in the other mode during the recognition process. Due to this capacity for mutual disambiguation, the performance of each error-prone mode potentially can be stabilized by the alternate mode whenever challenging usage conditions arise.

2.2.1.1 Accented Speaker Study. In a recent study, eight native speakers of English and eight accented speakers who represented different native languages (e.g., Mandarin, Tamil, Spanish, Turkish, Yoruba) each communicated 100 commands multimodally to the QuickSet system while using a hand-held PC. Sections 1.1 and 1.2 described the basic QuickSet system, and Fig. 2 illustrates its interface. With QuickSet, all participants could use multimodal speech and pen input to complete map-based simulation exercises. During testing, users accomplished a variety of tasks such as adding objects to a map (e.g., "Backburn zone" (draws irregular rectangular area)), moving objects (e.g., "Jeep follow this route"

FIG. 2. Diverse speakers completing commands multimodally using speech and gesture, which often would fail for a speech system due to varied accents.

(draws line)), and so forth. Details of the QuickSet system's signal and language processing, integration methods, and symbolic/statistical hybrid architecture have been summarized elsewhere [2,6,24].

In this study, data were collected on over 2000 multimodal commands, and the system's performance was analyzed for the overall multimodal recognition rate, recognition errors occurring within each system component (i.e., speech versus gesture recognition), and the rate of mutual disambiguation between speech and pen input during the integration process. When examining the rate of mutual disambiguation, all cases were assessed in which one or both recognizers failed to determine the correct lexical interpretation of the users' input, although the correct choice effectively was "retrieved" from lower down on an individual recognizer's n-best list to produce a correct final multimodal interpretation. The rate of mutual disambiguation per subject (MD_j) was calculated as the percentage of all their scorable integrated commands (N_j) in which the rank of the correct lexical choice on the multimodal n-best list (R_i^{MM}) was lower than the average rank of the correct lexical choice on the speech and gesture n-best lists (R_i^{s} and R_i^{g}), minus the number of commands in which the rank of the correct choice on the multimodal n-best list was higher than its average rank on the speech and gesture n-best lists, or

$$\mathrm{MD}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathrm{Sign}\left(\frac{R_i^s + R_i^g}{2} - R_i^{MM} \right)$$

MD was calculated both at the signal processing level (i.e., based on rankings in the speech and gesture signal n-best lists), and at the parse level after natural language processing (i.e., based on the spoken and gestural parse n-best lists). Scorable commands included all those that the system integrated successfully, and that contained the correct lexical information somewhere in the speech, gesture, and multimodal n-best lists. All significant MD results reported in this section [2.2.1] replicated across both signal and parse-level MD.

The results of this study confirmed that a multimodal architecture can support significant levels of mutual disambiguation, with one in eight user commands recognized correctly due to mutual disambiguation. Table Ia confirms that the speech recognition rate was much poorer for accented speakers (−9.5%), as would be expected, although their gesture recognition rate averaged slightly but significantly better (+3.4%). Table Ib reveals that the rate of mutual disambiguation (MD) was significantly higher for accented speakers (+15%) than for native speakers of English (+8.5%)—by a substantial 76%. As a result, Table Ia shows that the final multimodal recognition rate for accented speakers no longer differed significantly from the performance of native speakers. The main factor responsible

TABLE Ia

DIFFERENCE IN RECOGNITION RATE PERFORMANCE OF ACCENTED SPEAKERS, COMPARED WITH NATIVE ONES, DURING SPEECH, GESTURE, AND MULTIMODAL PROCESSING

Type of language processing	% Performance difference for accented speakers
Speech	−9.5*
Gesture	+3.4*
Multimodal	—

* Significant difference present.

TABLE Ib

MUTUAL DISAMBIGUATION (MD) RATE AND RATIO OF MD INVOLVING SPEECH SIGNAL PULL-UPS FOR NATIVE AND ACCENTED SPEAKERS

Type of MD metric	Native speakers	Accented speakers
Signal MD rate	8.5%	15.0% *
Ratio of speech pull-ups	0.35	0.65 *

* Significant difference present.

for closing this performance gap between groups was the higher rate of mutual disambiguation for accented speakers. Overall, a 41% reduction was revealed in the total error rate for spoken language processed within the multimodal architecture, compared with spoken language processed as a stand-alone [53].

Table Ib also reveals that speech recognition was the more fragile mode for accented speakers, with two-thirds of all mutual disambiguation involving pull-ups of their failed speech signals. However, the reverse was true for native speakers, with two-thirds of the mutual disambiguation in their case involving retrieval of failed ambiguous gesture signals. These data emphasize that there often are asymmetries during multimodal processing as to which input mode is more fragile in terms of reliable recognition. When one mode is expected to be less reliable, as is speech for accented speakers or during noise, then the most strategic multimodal design approach is to supplement the error-prone mode with an alternative one that can act as a natural complement and stabilizer by promoting mutual disambiguation.

Table II reveals that although single-syllable words represented just 40% of users' multimodal commands in these data, they nonetheless accounted for 58.2% of speech recognition errors. Basically, these brief monosyllabic commands were especially error prone because of the minimal amount of acoustic signal information available for the speech recognizer to process. These relatively fragile monosyllabic commands also accounted for 84.6% of the cases in which a failed speech interpretation was pulled up during the mutual disambiguation process, which was significantly greater than the rate observed for multisyllabic utterances [53].

2.2.1.2 Mobile Study. In a second study, 22 users interacted multimodally using the QuickSet system on a hand-held PC. Each user completed half of 100 commands in a quiet room (42 dB) while stationary, and the other half while mobile in a moderately noisy natural setting (40–60 dB), as illustrated in Fig. 3.

TABLE II

RELATION BETWEEN SPOKEN COMMAND LENGTH, THE PRESENCE OF SPEECH RECOGNITION ERRORS, AND THE PERCENTAGE OF MULTIMODAL COMMANDS WITH MUTUAL DISAMBIGUATION (MD) INVOLVING A SPEECH SIGNAL PULL-UP

	% Total commands in corpus	% Speech recognition errors	% MD with speech pull-ups
1 syllable	40	58.2	84.6*
2–7 syllables	60	41.8	15.4

* Significant difference present between monosyllabic and multisyllabic commands.

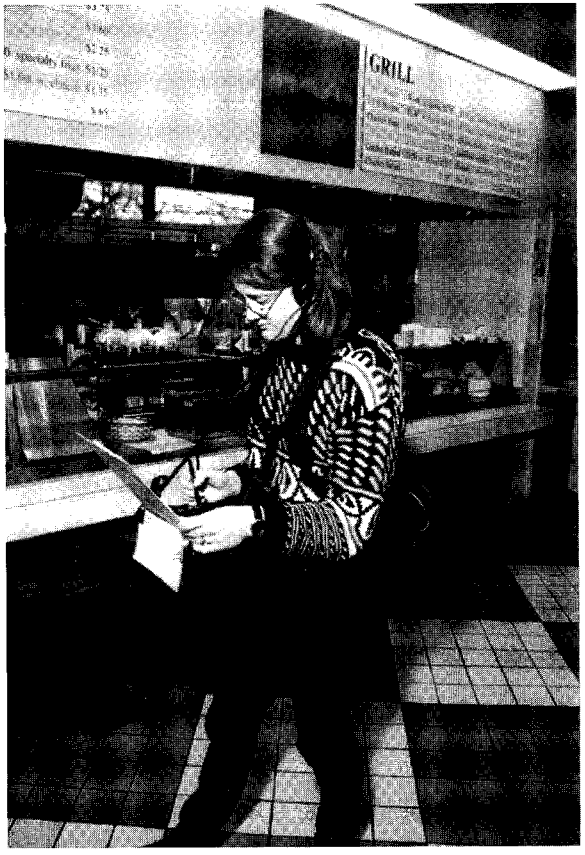

FIG. 3. Mobile user with a hand-held PC in a moderately noisy cafeteria, who is completing commands multimodally that often fail for a speech system.

Testing was replicated across microphones representing opposite quality, including a high-quality, close-talking, noise-canceling microphone, and also a low-quality, built-in microphone without noise cancellation. Over 2600 multimodal utterances were evaluated for the multimodal recognition rate, recognition errors occurring within each component recognizer, and the rate of mutual disambiguation between signals.

The results indicated that one in seven utterances were recognized correctly because of mutual disambiguation occurring during multimodal processing, even though one or both of the component recognizers failed to interpret the user's intended meaning. Table IIIa shows that the speech recognition rate was degraded

TABLE IIIa

DIFFERENCE IN RECOGNITION RATE PERFORMANCE IN MOBILE ENVIRONMENT, COMPARED WITH STATIONARY ONE, FOR SPEECH, GESTURE, AND MULTIMODAL PROCESSING

Type of language processing	% Performance difference when mobile
Speech	−10.0 *
Gesture	—
Multimodal	−8.0 *

* Significant difference present.

when speakers were mobile in a moderately noisy environment, compared with when they were stationary in a quiet setting (−10%). However, their gesture recognition rate did not decline significantly during mobility, perhaps because pen input involved brief one- to three-stroke gestures. Table IIIb reveals that the rate of mutual disambiguation in the mobile condition (+16%) also averaged substantially higher than the same user's stationary rate (+9.5%). As a result, Table IIIa confirms a significant narrowing of the gap between mobile and stationary recognition rates (to −8.0%) during multimodal processing, compared with spoken language processing alone. In fact, 19–35% relative reductions in the total error rate (for noise-canceling versus built-in microphones, respectively) were observed when speech was processed within the multimodal architecture [58]. Finally, the general pattern of results obtained in this mobile study replicated across opposite types of microphone technology.

When systems must process speech in natural contexts that involve variable levels of noise, and qualitatively different types of noise (e.g., abrupt onset, phase-in/phase-out), the problem of supporting robust recognition is extremely difficult. 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 that users make) simply is a challenging problem. As a result, mobile speech processing remains an unsolved problem for traditional speech recognition. In the face of such challenges, a multimodal architecture that supports mutual disambiguation potentially can provide

TABLE IIIb

MUTUAL DISAMBIGUATION (MD) RATE AND RATIO OF MD INVOLVING SPEECH SIGNAL PULL-UPS IN STATIONARY AND MOBILE ENVIRONMENTS

Type of MD metric	Stationary	Mobile
Signal MD rate	9.5%	16.0% *
Ratio of speech pull-ups	.26	.34 *

* Significant difference present.

greater stability and a more viable long-term avenue for managing errors in emerging mobile interfaces. This theme also is central to the performance advantages identified for multimodal speech and lip movement systems, which are described in Section 2.2.2.

One unique aspect of this mobile study was its focus on testing during actual use of an implemented multimodal system while users were mobile in a natural field environment. Such performance testing was possible because of the state of development of multimodal speech and pen systems, which now are beginning to transition into commercial applications. It also was possible because of the emerging research infrastructure now becoming available for collecting mobile field data [58]. In addition, this mobile study was unique in its examination of performance during naturalistic noisy conditions, especially the inclusion of nonstationary noise. As a result, the present data provide information on the expected performance advantages of multimodal systems in moderately noisy field settings, with implications for the real-world commercialization of new mobile interfaces.

In summary, in both of the studies described in this section, even though one or both of the component recognizers failed to identify users' intended meaning, the architectural constraints imposed by the multimodal system's unification process ruled out incompatible speech and pen signal integrations. These unification constraints effectively pruned recognition errors from the *n*-best lists of the component recognizers, which resulted in the retrieval of correct lexical information from lower down on their lists, producing a correct final multimodal interpretation. This process suppressed many errors that would have occurred, such that users never experienced them. It also had an especially large impact on reducing the speech recognition errors that otherwise were so prevalent for accented speakers and in noisy environments.

2.2.2 *Robustness of Multimodal Speech and Lip Movement Systems*

During the natural audio–visual perception of speech, human listeners typically observe a speaker's lip and facial movements while attending to speech. Furthermore, their accurate interpretation of speech is well known to be superior during multimodal speech perception, compared with acoustic-only speech processing [5,30,32,34]. In noisy environments, which include most natural field environments, visual information about a speaker's lip movements can be particularly valuable for the accurate interpretation of speech. However, there also are large individual and cultural differences in the information available in visible lip movements, as well as in people's ability and tendency to lip-read [7]. For example, the hearing impaired, elderly, and nonnative speakers all typically rely

more heavily on visual lip movements when they attend to speech, so for these populations accurate interpretation can depend critically on combined audio–visual processing [100,101]. The cognitive science literature generally has provided a good foundation for understanding many aspects of the design and expected value of multimodal speech and lip movement systems.

In many of the multimodal speech and lip movement systems developed during the 1980s and 1990s, error suppression also has been observed [3,37,39,43,45,61–65,102]. This literature has investigated the use of visually derived information about a speaker's lip movements (visemes) to improve recognition of acoustic speech (phonemes). The primary focus of this research has been on demonstrating robustness improvement during the audio–visual processing of speech *during noise*, compared with acoustic-only speech processing, with demonstrations of a larger boost in robustness as the noise level increases and speech recognition errors rise. Robustness improvements for multimodal speech and lip movement systems that have been reported *under noise-free conditions* actually have been relatively small when they occur at all, typically with less than a 10% relative error reduction [102]. In fact, sometimes a performance penalty occurs during the audio–visual processing of noise-free speech, largely as a consequence of adopting approaches designed to handle speech in noise [103]. On the other hand, robustness improvements of over 50% relative error reduction frequently have been documented under noisy conditions [39,45,46,61,65,102].

2.2.2.1 Profile of Typical Study. In typical studies exploring performance enhancement in multimodal speech and lip movement systems, researchers have compared different approaches for audio-only, visual-only, and audio–visual system processing. Typically, testing has been done on a limited single-speaker corpus involving read materials such as nonsense words or digits [39,43]. Artificial stationary noise (e.g., white noise) then is added to generate conditions representing a range of different signal-to-noise ratio (SNR) decibel levels, for example, graduated intervals between −5 and +25 dB. Most assessments have been performed on isolated-word speaker-dependent speech systems [39], although more recent studies now are beginning to examine continuous speech recognition as well [45]. The most common goal in these studies has been a basic demonstration of whether word error rates for audio–visual speech processing exceed those for audio-only and video-only processing, preferably at all levels of additive noise. Frequently, different integration strategies for audio–visual processing also are compared in detail. As described previously, the most common result has been to find the largest enhancements of audio–visual performance at the most degraded noise levels, and modest or no enhancement in a noise-free context.

Unlike studies on multimodal speech and pen systems, research on the performance of multimodal speech and lip movement systems has not focused on the

mutual disambiguation of information that can occur between two rich input modes, but rather on the bootstrapping of speech recognition under noisy conditions. In addition, studies conducted in this area have not involved testing with fully implemented systems in actual noisy field settings. They likewise have been limited to testing on stationary noise sources (for a recent exception, see DuPont and Luettin's research [45]), rather than the more realistic and challenging nonstationary sources common during mobile use. Future research in this area will need to include more realistic test conditions before results can be generalized to situations of commercial import. More recent research in this area now is beginning to train systems and evaluate multimodal integration techniques on increasingly large multiparty corpora, and also to develop multimodal audio–visual systems for a variety of potential applications (e.g., speech recognition, speaker recognition, speech event detection) [3,45].

Currently, researchers in this area are striving to develop new integration techniques that can support general robustness advantages across the spectrum of noise conditions, from extremely adverse environments with SNR ranging 0 to −22 dB, to quiet environments with SNR ranging 20–30 dB. One goal is to develop multimodal integration techniques that yield generally superior robustness in widely varied and potentially changing environment conditions, such as those anticipated during mobile use. A second goal is to demonstrate larger improvements for audio–visual processing over audio-only in noise-free environments, which has been relatively elusive to date [104]. Late-integration fusion (i.e., "decision-level") and hybrid integration techniques, such as those used in multimodal speech and pen systems, generally have become viewed as good avenues for achieving these robustness goals [3,43,45,62,65,102].

Recent work also has begun to focus on audio–visual robustness gains achievable through adaptive processing, in particular various techniques for stream weight estimation [45,63,64]. For example, a recent experiment by Potamianos and Neti [64] of IBM–Watson reported over a 20% relative error reduction based on an n-best stream likelihood dispersion measure. Further work on adaptive multimodal processing is an important research direction in need of additional attention. Issues as basic as determining the key criteria and strategies needed to accomplish intelligent adaptation in natural field settings still are very poorly understood. In general, early attempts to adapt multimodal audio–visual processing based on simple engineering concepts will need to be superceded by empirically validated strategies. For example, automated dynamic weighting of the audio and visual input modes as a function of SNR estimates [46,47] is known to be problematic because it fails to take into account the impact of users' Lombard adaptations (for discussion, see Oviatt's research [105]).

Like the literature on multimodal speech and pen interaction, research in this area has identified key complementarities between the audio speech signal and corresponding visible speech movements [29,33,106]. For example, place of articulation is difficult to discriminate auditorally for consonants, but easy to distinguish visually from the position of the teeth, tongue, and lips. Natural feature-level complementarities also have been identified between visemes and phonemes for vowel articulation, with vowel rounding better conveyed visually, and vowel height and backness better revealed auditorally [29,33].

Some speech and lip movement systems have developed heuristic rules incorporating information about the relative confusability of different kinds of phonemes within their audio and visual processing components [107]. Future systems that incorporate phoneme-level information of this kind are considered a potentially promising avenue for improving robustness. In particular, research on the misclassification of consonants and vowels by audio–visual systems has emphasized the design recommendation that the visual component be weighted more heavily when discriminating place and manner of articulation, but less heavily when determining voicing [65]. Research by Silsbee and colleagues [65] has indicated that when consonant versus vowel classification tasks are considered separately, although no robustness enhancement occurs for audio–visual processing of consonants during noise-free conditions, an impressive 61% relative error reduction is obtained for vowels [65]. These results underscore the potential value of applying cognitive science findings to the design of future adaptive systems.

Finally, like the literature on multimodal speech and pen systems, in this research area brief spoken monosyllables have been associated with larger magnitude robustness gains during audio–visual processing, compared to multisyllabic utterances [108]. This is largely because monosyllables contain relatively impoverished acoustic information, and therefore are subject to higher rates of speech recognition errors. This finding in the speech and lip movement literature basically is parallel to the higher rate of mutual disambiguation reported for monosyllables in the multimodal speech and pen literature [53]. As will be discussed in Section 2.3, this replicated finding suggests that monosyllables may represent one of the *targets of opportunity* for future multimodal system design.

2.3 Multimodal Design Strategies for Optimizing Robustness

From the emerging literature on multimodal system performance, especially the error suppression achievable with such systems, there are several key concepts that surface as important for their design. The following are examples of fertile

research strategies known to be relevant to improving the robustness of future multimodal systems:

- *Increase the number of input modes interpreted within the multimodal system.* This principle is effective because it supports effective supplementation and disambiguation of partial or conflicting information that may be present in any individual input mode. Current bimodal systems largely are successful due to their elementary fusion of information sources. However, according to this general principle, future multimodal systems could optimize robustness further by combining additional information sources—for example, three or more input modes. How much additional robustness gain can be expected as a function of incorporating additional sources of information is an issue that remains to be evaluated in future research.
- *Combine input modes that represent semantically rich information sources.* In order to design multimodal systems that support mutual disambiguation, a minimum of two semantically rich input modes is required. Both types of multimodal system discussed in this chapter process two semantically rich input modes, and both have demonstrated enhanced error suppression compared with unimodal processing. In contrast, multimodal systems that combine only one semantically rich input mode (e.g., speech) with a second that is limited in information content (e.g., mouse, touch, or pen input only for selection) cannot support mutual disambiguation. However, even these more primitive multimodal systems can support disambiguation of the rich input mode to some degree by the more limited one. For example, when pointing to select an interface entity or input field, the natural language processing can be constrained to a reduced set of viable interpretations, thereby improving the accuracy of spoken language recognition [109].
- *Increase the heterogeneity of input modes combined within the multimodal system.* In order to bootstrap the joint potential of two input modes for collecting the relevant information needed to achieve mutual disambiguation of partial or conflicting information during fusion, one strategy is to sample from a broad range of qualitatively different information sources. In the near term, the most likely candidates for new modes to incorporate within multimodal systems involve vision-based recognition technologies. Specific goals and strategies for achieving increased heterogeneity of information, and how successfully they may optimize overall multimodal system robustness, is a topic that needs to be addressed in future research. One specific strategy for achieving heterogeneity is described in the next section.
- *Integrate maximally complementary input modes.* One goal in the design of multimodal systems is to combine modes into a well-integrated system. If

designed opportunistically, such a system should integrate complementary modalities to yield a highly synergistic blend in which the strengths of each mode can be capitalized upon and used to overcome weaknesses in the other [11]. As discussed earlier, in the multimodal speech and lip movement literature, natural feature-level complementarities already have been identified between visemes and phonemes [29,33]. In multimodal speech and pen research, the main complementarity involves visual–spatial semantic content. Whereas visual–spatial information is uniquely and clearly indicated via pen input, the strong descriptive capabilities of speech are better suited for specifying temporal and other nonspatial information [56,110].

In general, this design approach promotes the philosophy of using modalities to their natural advantage, and it also represents a strategy for combining modes in a manner that can generate mutual disambiguation. In fact, achieving multimodal performance gains of the type described earlier in this chapter is well known to depend in part on successful identification of the unique semantic complementarities of a given pair of input modes. As discussed in Section 2.2.1, when one mode is expected to be less reliable (e.g., speech for accented speakers or during noise), then the most strategic multimodal design approach is to supplement the error-prone mode with a second one that can act as a natural complement and stabilizer in promoting mutual disambiguation. Future research needs to explore asymmetries in the reliability of different input modes, as well as the main complementarities that exist between modes that can be leveraged during multimodal system design.

- *Develop multimodal processing techniques that retain information.* In addition to the general design strategies outlined above, it also is important to develop multimodal signal processing, language processing, and architectural techniques that retain information and make it available during decision-level fusion. For example, alternative interpretations should not be pruned prematurely from each of the component recognizers' *n*-best lists. Excessive pruning of *n*-best list alternatives (i.e., by setting probability estimate thresholds too high) could result in eliminating the information needed for mutual disambiguation to occur. This is because the correct partial information must be present on each recognizer's *n*-best list in order for the correct final multimodal interpretation to be formed during unification.

The following are research strategies that are known to be relevant for successfully applying multimodal system design to *targets of opportunity* in which the

greatest enhancement of robustness is likely to be demonstrated over unimodal system design:

- *Apply multimodal system design to brief information segments for which robust recognition is known to be unreliable.* As outlined in Sections 2.2.1 and 2.2.2, brief segments of information are the most fragile and subject to error during recognition (e.g., monosyllabic acoustic content during speech recognition). They also are selectively improved during multimodal processing in which additional information sources are used to supplement interpretation.
- *Apply multimodal system design to challenging user groups and usage environments for which robust recognition is known to be unreliable.* When a recognition-based component technology is known to be selectively faulty for a given user group or usage environment, then a multimodal interface can be used to stabilize errors and improve the system's average recognition accuracy. As discussed earlier, accented speakers and noisy mobile environments are more prone to precipitate speech recognition errors. In such cases, a multimodal interface that processes additional information sources can be crucial in disambiguating the error-prone speech signal, sometimes recovering performance to levels that match the accuracy of nonrisk conditions. Further research needs to continue investigating other potential targets of opportunity that may benefit selectively from multimodal processing, including especially complex task applications, error-prone input devices (e.g., laser pointers), and so forth.

In discussing the above strategies, there is a central theme that emerges. Whenever information is too scant or ambiguous to support accurate recognition, a multimodal interface can provide an especially opportune solution to fortify robustness. Furthermore, the key design strategies that contribute to the enhanced robustness of multimodal interfaces are those that add greater breadth and richness to the information sources that are integrated within a given multimodal system. Essentially, the broader the information collection net cast, the greater the likelihood missing or conflicting information will be resolved, leading to successful disambiguation of user input during the recognition process.

2.4 Performance Metrics as Forcing Functions for Robustness

In the past, the speech community has relied almost exclusively on the assessment of WER to calibrate the performance accuracy of spoken language systems. This metric has served as the basic forcing function for comparing and iterating

spoken language systems. In particular, WER was used throughout the DARPA-funded Speech Grand Challenge research program [78] to compare speech systems at various funded sites. Toward the end of this research program, it was widely acknowledged that although metrics are needed as a forcing function, nonetheless reliance on any single metric can be risky and counterproductive to the promotion of high-quality research and system building. This is because a singular focus on developing technology to meet the demands of any specific metric essentially encourages the research community to adopt a narrow and conservative set of design goals. It also tends to encourage relatively minor iterative algorithmic adaptations during research and system development, rather than a broader and potentially more productive search for innovative solutions to the hardest problems. When innovative or even radically different strategies are required to circumvent a difficult technical barrier, then new performance metrics can act as a stimulus and guide in advancing research in the new direction. Finally, in the case of the speech community's overreliance on WER, one specific adverse consequence was the general disincentive to address many essential user-centered design issues that could have reduced errors and improved error handling in spoken language systems.

During the development of multimodal systems, one focus of early assessments clearly has been on the demonstration of improved robustness over unimodal speech systems. To track this, researchers have calculated an overall multimodal recognition rate, although often summarized at the utterance level and with additional diagnostic information about the performance of the system's two component recognizers. This has provided a global assessment tool for indexing the average level of multimodal system accuracy, as well as the basic information needed for comparative analysis of multimodal versus unimodal system performance.

However, as an alternative approach to traditional speech processing, multimodal research also has begun to adopt new and more specialized metrics, such as a given system's rate of mutual disambiguation. This concept has been valuable for assessing the degree of error suppression achievable in multimodal systems. It also has provided a tool for assessing each input mode's ability to disambiguate errors in the other mode. This latter information has assisted in clarifying the relative stability of each mode, and also in establishing how effectively two modes work together to supply the complementary information needed to stabilize system performance. In this respect, the mutual disambiguation metric has significant *diagnostic capabilities* beyond simply summarizing the average level of system accuracy.

As part of exploratory research, the mutual disambiguation metric also is beginning to be used to define *in what circumstances* a particular input mode is effective

at stabilizing the performance of a more fragile mode. In this sense, it is playing an active role in exploring *user-centered design issues* relevant to the development of new multimodal systems. It also is elucidating the dynamics of error suppression. In the future, other new metrics that reflect concepts of central importance to the development of emerging multimodal systems will be needed.

3. Future Directions: Breaking the Robustness Barrier

The computer science community is just beginning to understand how to design innovative, well-integrated, and robust multimodal systems. To date, most multimodal systems remain bimodal, and recognition technologies related to several human senses (e.g., haptics, smell) have yet to be well represented within multimodal interfaces. As with past multimodal systems, the design and development of new types of multimodal system that include such modes will not be achievable through intuition alone. Rather, it will depend on knowledge of the usage and natural integration patterns that typify people's combined use of various input modes. This means that the successful design of new multimodal systems will continue to require guidance from cognitive science on the coordinated human perception and production of natural modalities. In this respect, multimodal systems only can flourish through multidisciplinary cooperation and teamwork among those working on different component technologies. The multimodal research community also could benefit from far more cross-fertilization among researchers representing the main subareas of multimodal expertise, especially those working in the more active areas of speech and pen and speech and lip movement research. Finally, with multimodal research projects and funding expanding in Europe, Japan, and elsewhere, the time is ripe for more international collaboration in this research area.

To achieve commercialization and widespread dissemination of multimodal interfaces, more general, robust, and scalable multimodal architectures will be needed, which now are beginning to emerge. Most multimodal systems have been built during the past decade, and they are research-level systems. However, in several cases they now have developed beyond the prototype stage, and are being integrated with other software at academic and federal sites, or are beginning to appear as newly shipped products [2,19]. Future research will need to focus on developing hybrid symbolic/statistical architectures based on large corpora and refined fusion techniques in order to optimize multimodal system robustness. Research also will need to develop new architectures capable of flexibly coordinating numerous multimodal–multisensor system components to support new directions in adaptive processing. To transcend the

robustness barrier, research likewise will need to explore new natural language, dialogue processing, and statistical techniques for optimizing mutual disambiguation among the input modes combined within new classes of a multimodal system.

As multimodal interfaces gradually progress toward supporting more robust and human-like perception of users' natural activities in context, they will need to expand beyond rudimentary bimodal systems to ones that incorporate three or more input modes. Like biological systems, they should be generalized to include input from qualitatively different and semantically rich information sources. This increase in the number and heterogeneity of input modes can effectively broaden the reach of advanced multimodal systems, and provide them with access to the discriminative information needed to reliably recognize and process users' language, actions, and intentions in a wide array of different situations. Advances of this kind are expected to contribute to a new level of robustness or *hybrid vigor* in multimodal system performance. This trend already has been initiated within the field of biometrics research, which is combining recognition of multiple behavioral modes with physiological ones to achieve reliable person identification and verification under challenging field conditions. To support increasingly pervasive multimodal interfaces, these combined information sources ideally must include data collected from a wide array of sensors as well as input modes, and from both active and passive forms of user input.

Very few existing multimodal systems that involve speech recognition currently include any adaptive processing. With respect to societal impact, the shift toward adaptive multimodal interfaces is expected to provide significantly enhanced usability for a diverse range of everyday users, including young and old, experienced and inexperienced, able-bodied and disabled. Such interfaces also will be far more personalized and appropriately responsive to the changing contexts induced by mobility than interfaces of the past. With respect to robustness, adaptivity to the user, ongoing task, dialogue, environmental context, and input modes will collectively generate constraints that can greatly improve system reliability.

In the future, adaptive multimodal systems will require active tracking of potentially discriminative information, as well as flexible incorporation of additional information sources during the process of fusion and interpretation. In this respect, future multimodal interfaces and architectures will need to be able to engage in flexible reconfiguration, such that specific types of information can be integrated as needed when adverse conditions arise (e.g., noise), or if the confidence estimate for a given interpretation falls too low. The successful design of future adaptive multimodal systems could benefit from a thoughtful examination of the models already provided by biology and cognitive science on intelligent adaptation during perception, as well as from the literature on robotics.

4. Conclusion

In summary, a well-designed multimodal system that fuses two or more information sources can be an effective means of reducing recognition uncertainty. Performance advantages have been demonstrated for different modality combinations (speech and pen, speech and lip movements), as well as for varied tasks and different environments. Furthermore, the average error suppression achievable with a multimodal system, compared with a unimodal spoken language one, can be very substantial. These findings indicate that promising but error-prone recognition-based technologies are increasingly likely to be embedded within multimodal systems in order to achieve commercial viability during the next decade.

Recent research also has demonstrated that multimodal systems can perform more stably for challenging real-world user groups and usage contexts. For this reason, they are expected to play an especially central role in the emergence of mobile interfaces, and in the design of interfaces for every-person universal access. In the long term, adaptive multimodal–multisensor interfaces are viewed as a key avenue for supporting far more pervasive interfaces with entirely new functionality not supported by computing of the past.

ACKNOWLEDGMENTS

I thank the National Science Foundation for their support over the past decade, which has enabled me to pursue basic exploratory research on many aspects of multimodal interaction, interface design, and system development. The preparation of this chapter has been supported by NSF Grant IRI-9530666 and NSF Special Extension for Creativity (SEC) Grant IIS-9530666. This work also has been supported by Contracts DABT63-95-C-007 and N66001-99-D-8503 from DARPA's Information Technology and Information Systems Office, and Grant N00014-99-1-0377 from ONR. I also thank Phil Cohen and others in the Center for Human–Computer Communication for many insightful discussions, and Dana Director, Rachel Coulston, and Kim Tice for expert assistance with manuscript preparation.

REFERENCES

[1] Benoit, C., Martin, J. C., Pelachaud, C., Schomaker, L., and Suhm, B. (2000). "Audio-visual and multimodal speech-based systems." *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation* (D. Gibbon, I. Mertins, and R. Moore, Eds.), pp. 102–203. Kluwer Academic, Boston.

[2] Oviatt, S. L., Cohen, P. R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., and Ferro, D. (2000). "Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions." *Human Computer Interaction*, **15**, 4, 263–322.

[Reprinted in *Human-Computer Interaction in the New Millennium* (J. Carroll, Ed.), Chap. 19, pp. 421–456. Addison–Wesley, Reading, MA, 2001.]

[3] 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." *Proceedings of the International Conference on Spoken Language Processing*, Beijing, **3**, 11–14.

[4] Pankanti, S., Bolle, R. M., and Jain, A. (Eds.) (2000). "Biometrics: The future of identification." *Computer*, **33**, 2, 46–80.

[5] Benoit, C., and Le Goff, B. (1998). "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP." *Speech Communication*, **26**, 117–129.

[6] Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. (1997). "Quickset: Multimodal interaction for distributed applications." *Proceedings of the Fifth ACM International Multimedia Conference*, pp. 31–40. ACM Press, New York.

[7] Stork, D. G., and Hennecke, M. E. (Eds.) (1996). *Speechreading by Humans and Machines*. Springer-Verlag, New York.

[8] Turk, M., and Robertson, G. (Eds.) (2000). "Perceptual user interfaces." *Communications of the ACM* (special issue on Perceptual User Interface), **43**, 3, 32–70.

[9] Zhai, S., Morimoto, C., and Ihde, S. (1999). "Manual and gaze input cascaded (MAGIC) pointing." *Proceedings of the Conference on Human Factors in Computing Systems (CHI'99)*, pp. 246–253. ACM Press, New York.

[10] Bolt, R. A. (1980). "Put-that-there: Voice and gesture at the graphics interface." *Computer Graphics*, **14**, 3, 262–270.

[11] Cohen, P. R., Dalrymple, M., Moran, D. B., Pereira, F. C. N., Sullivan, J. W., Gargan, R. A., Schlossberg, J. L., and Tyler, S. W. (1989). "Synergistic use of direct manipulation and natural language." *Proceedings of the Conference on Human Factors in Computing Systems (CHI'89)*, pp. 227–234. ACM Press, New York. [Reprinted in *Readings in Intelligent User Interfaces* (Maybury and Wahlster, Eds.), pp. 29–37, Morgan Kaufmann, San Francisco.]

[12] Kobsa, A., Allgayer, J., Reddig, C., Reithinger, N., Schmauks, D., Harbusch, K., and Wahlster, W. (1986). "Combining deictic gestures and natural language for referent identification." *Proceedings of the 11th International Conf. on Computational Linguistics*, Bonn, Germany, pp. 356–361.

[13] Neal, J. G., and Shapiro, S. C. (1991). "Intelligent multimedia interface technology." *Intelligent User Interfaces* (J. W. Sullivan and S. W. Tyler, Eds.), pp. 11–43. ACM Press, New York.

[14] Seneff, S., Goddeau, D., Pao, C., and Polifroni, J. (1996). "Multimodal discourse modeling in a multi-user multi-domain environment." *Proceedings of the International Conference on Spoken Language Processing* (T. Bunnell and W. Idsardi, Eds.), Vol. 1, pp. 192–195. University of Delaware and A. I. duPont Institute.

[15] Siroux, J., Guyomard, M., Multon, F., and Remondeau, C. (1995). "Modeling and processing of the oral and tactile activities in the Georal tactile system."

Proceedings of the International Conference on Cooperative Multimodal Communication, Theory & Applications. Eindhoven, Netherlands.

[16] Wahlster, W. (1991). "User and discourse models for multimodal communciation." *Intelligent User Interfaces* (J. W. Sullivan and S. W. Tyler, Eds.), Chap. 3, pp. 45–67. ACM Press, New York.

[17] Oviatt, S. L., and Cohen, P. R. (2000). "Multimodal systems that process what comes naturally." *Communications of the ACM*, **43**, 3, 45–53.

[18] Rubin, P., Vatikiotis-Bateson, E., and Benoit, C. (Eds.) (1998). *Speech Communication* (special issue on audio-visual speech processing). **26**, 1–2.

[19] Oviatt, S. L. (2002). "Multimodal Interfaces." *Handbook of Human-Computer Interaction* (J. Jacko and A. Sears, Eds.). Lawrence Erlbaum, Mahwah, NJ.

[20] Oviatt, S. L., Cohen, P. R., Fong, M. W., and Frank, M. P. (1992). "A rapid semi-automatic simulation technique for investigating interactive speech and handwriting." *Proceedings of the International Conference on Spoken Language Processing*, University of Alberta, Vol. 2, pp. 1351–1354.

[21] Bers, J., Miller, S., and Makhoul, J. (1998). "Designing conversational interfaces with multimodal interaction." *DARPA Workshop on Broadcast News Understanding Systems*, pp. 319–321.

[22] Cheyer, A. (1998). "MVIEWS: Multimodal tools for the video analyst." *Proceedings of the International Conference on Intelligent User Interfaces (IUI'98)*, pp. 55–62. ACM Press, New York.

[23] Waibel, A., Suhm, B., Vo, M. T., and Yang, J. (1997). "Multimodal interfaces for multimedia information agents." *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE-ICASSP)*, Vol. 1, pp. 167–170. IEEE Press, Menlo Park, CA.

[24] Wu, L., Oviatt, S., and Cohen, P. (1999). "Multimodal integration: A statistical view." *IEEE Transactions on Multimedia*, **1**, 4, 334–342.

[25] Bangalore, S., and Johnston, M. (2000). "Integrating multimodal language processing with speech recognition." *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000)* (B. Yuan, T. Huang, and X. Tang, Eds.), Vol. 2, pp. 126–129. Chinese Friendship, Beijing.

[26] Denecke, M., and Yang, J. (2000). "Partial information in multimodal dialogue." *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000)* (B. Yuan, T. Huang, and X. Tang, Eds.), pp. 624–633. Chinese Friendship, Beijing.

[27] Bernstein, L., and Benoit, C. (1996). "For speech perception by humans or machines, three senses are better than one." *Proceedings of the International Conference on Spoken Language Processing*, **3**, 1477–1480.

[28] Cohen, M. M., and Massaro, D. W. (1993). "Modeling coarticulation in synthetic visible speech." *Models and Techniques in Computer Animation* (N. M. Thalmann and D. Thalmann, Eds.), pp. 139–156. Springer-Verlag, Berlin.

[29] Massaro, D. W., and Stork, D. G. (1998). "Sensory integration and speechreading by humans and machines." *American Scientist*, **86**, 236–244.

[30] McGrath, M., and Summerfield, Q. (1985). "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults." *Journal of the Acoustical Society of America*, **77**, 2, 678–685.

[31] McGurk, H., and MacDonald, J. (1976). "Hearing lips and seeing voices." *Nature*, **264**, 746–748.

[32] McLeod, A., and Summerfield, Q. (1987). "Quantifying the contribution of vision to speech perception in noise." *British Journal of Audiology*, **21**, 131–141.

[33] Robert-Ribes, J., Schwartz, J. L., Lallouache, T., and Escudier, P. (1998). "Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise." *Journal of the Acoustical Society of America*, **103**, 6, 3677–3689.

[34] Sumby, W. H., and Pollack, I. (1954). "Visual contribution to speech intelligibility in noise." *Journal of the Acoustical Society of America*, **26**, 212–215.

[35] Summerfield, A. Q. (1992). "Lipreading and audio-visual speech perception." *Philosophical Transactions of the Royal Society of London, Series B*, **335**, 71–78.

[36] Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., and Terzopoulos, D. (1996). "The dynamics of audiovisual behavior in speech." *Speechreading by Humans and Machines: Models, Systems and Applications* (D. G. Stork and M. E. Hennecke, Eds.), NATO ASI Series, Series F: Computer and Systems Sciences 150, pp. 221–232. Springer-Verlag, Berlin.

[37] Petajan, E. D. (1984). *Automatic Lipreading to Enhance Speech Recognition*, Ph.D. thesis. University of Illinois at Urbana–Champaign.

[38] Brooke, N. M., and Petajan, E. D. (1986). "Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics." *Proceedings of the International Conference on Speech Input and Output: Techniques and Applications*, **258**, 104–109.

[39] Adjoudani, A., and Benoit, C. (1995). "Audio-visual speech recognition compared across two architectures." *Proceedings of the Eurospeech Conference*, Madrid, Spain, Vol. 2, pp. 1563–1566.

[40] Bregler, C., and Konig, Y. (1994). "Eigenlips for robust speech recognition." *Proceedings of the International Conference on Acoustics Speech and Signal Processing (IEEE-ICASSP)*, Vol. 2, pp. 669–672.

[41] Goldschen, A. J. (1993). *Continuous Automatic Speech Recognition by Lipreading*, Ph.D. thesis. Department of Electrical Engineering and Computer Science, George Washington University.

[42] Silsbee, P. L., and Su, Q. (1996). "Audiovisual sensory integration using Hidden Markov Models." *Speechreading by Humans and Machines: Models, Systems and Applications* (D. G. Stork and M. E. Hennecke, Eds.), NATO ASI Series, Series F: Computer and Systems Sciences 150, pp. 489–504. Springer-Verlag, Berlin.

[43] Tomlinson, M. J., Russell, M. J., and Brooke, N. M. (1996). "Integrating audio and visual information to provide highly robust speech recognition." *Proceedings of the International Conference on Acoustics Speech and Signal Processing (IEEE-ICASSP)*, pp. 821–824.

[44] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.) (2000). *Embodied conversational agents*. MIT Press, Cambridge, MA.

[45] Dupont, S., and Luettin, J. (2000). "Audio-visual speech modeling for continuous speech recognition." *IEEE Transactions on Multimedia*, **2**, 3, 141–151.

[46] Meier, U., Hürst, W., and Duchnowski, P. (1996). "Adaptive bimodal sensor fusion for automatic speechreading." *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE-ICASSP)*, pp. 833–836. IEEE Press, Menlo Park, CA.

[47] Rogozan, A., and Deglise, P. (1998). "Adaptive fusion of acoustic and visual sources for automatic speech recognition." *Speech Communication*, **26**, 1–2, 149–161.

[48] Choudhury, T., Clarkson, B., Jebara, T., and Pentland, S. (1999). "Multimodal person recognition using unconstrained audio and video." *Proceedings of the 2nd International Conference on Audio-and-Video-based Biometric Person Authentication*, Washington, DC, pp. 176–181.

[49] Lee, J. (2001). "Retooling products so all can use them." *New York Times*, June 21.

[50] Jorge, J., Heller, R., and Guedj, R. (Eds.) (2001). *Proceedings of the NSF/EC Workshop on Universal Accessibility and Ubiquitous Computing: Providing for the Elderly*, Alcácer do Sal, Portugal, 22–25 May. Available at `http://immi.inesc.pt/alcacer01/procs/papers-list.html`.

[51] Oviatt, S. L., and van Gent, R. (1996). "Error resolution during multimodal human-computer interaction." *Proceedings of the International Conference on Spoken Language Processing*, Vol. 2, pp. 204–207. University of Delaware Press.

[52] Oviatt, S. L., Bernard, J., and Levow, G. (1998). "Linguistic adaptation during error resolution with spoken and multimodal systems." *Language and Speech* (special issue on prosody and conversation), **41**, 3–4, 419–442.

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

[54] Rudnicky, A., and Hauptman, A. (1992). "Multimodal interactions in speech systems." *Multimedia Interface Design, Frontier Series* (M. Blattner and R. Dannenberg, Eds.), pp. 147–172. ACM Press, New York.

[55] Suhm, B. (1998). *Multimodal Interactive Error Recovery for Non-conversational Speech User Interfaces*, Ph.D. thesis. Karlsruhe University, Germany.

[56] Oviatt, S. L. (1997). "Multimodal interactive maps: Designing for human performance." *Human–Computer Interaction* (special issue on multimodal interfaces), **12**, 93–129.

[57] Oviatt, S. L., and Kuhn, K. (1998). "Referential features and linguistic indirection in multimodal language." *Proceedings of the International Conference on Spoken Language Processing*, ASSTA Inc., Sydney, Australia, Vol. 6, pp. 2339–2342.

[58] Oviatt, S. L. (2000). "Multimodal system processing in mobile environments." *Proceedings of the Thirteenth Annual ACM Symposium on User Interface Software Technology (UIST 2000)*, pp. 21–30. ACM Press, New York.

[59] Oviatt, S. L. (2000). "Taming recognition errors with a multimodal architecture." *Communications of the ACM* (special issue on conversational interfaces), **43**, 9, 45–51.

[60] Erber, N. P. (1975). "Auditory-visual perception of speech." *Journal of Speech and Hearing Disorders*, **40**, 481–492.

[61] Bregler, C., Omohundro, S. M., Shi, J., and Konig, Y. (1996). "Towards a robust speechreading dialog system." *Speechreading by Humans and Machines: Models, Systems and Applications* (D. G. Stork and M. E. Hennecke, Eds.), NATO ASI Series, Series F: Computer and Systems Sciences 150, pp. 409–423. Springer-Verlag, Berlin.

[62] Brooke, M. (1996). "Using the visual component in automatic speech recognition." *Proceedings of the International Conference on Spoken Language Processing*, Vol. 3, pp. 1656–1659.

[63] Nakamura, S., Ito, H., and Shikano, K. (2000). "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition." *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)* (B. Yuan, T. Huang and X. Tang, Eds.), Vol. 3, pp. 20–24. Chinese Friendship Publishers, Beijing.

[64] Potamianos, G., and Neti, C. (2000). "Stream confidence estimation for audio-visual speech recognition." *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)* (B. Yuan, T. Huang, and X. Tang, Eds.), Vol. 3, pp. 746–749. Chinese Friendship Publishers, Beijing.

[65] Silsbee, P. L., and Bovik, A. C. (1996). "Computer lipreading for improved accuracy in automatic speech recognition." *IEEE Transactions on Speech and Audio Processing*, **4**, 5, 337–351.

[66] Murphy, R. R. (1996). "Biological and cognitive foundations of intelligent sensor fusion." *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, **26**, 1, 42–51.

[67] Lee, D. (1978). "The functions of vision." *Modes of Perceiving and Processing Information* (H. L. Pick and E. Saltzman, Eds.), pp. 159–170. Wiley, New York.

[68] Pick, H. L., and Saltzman, E. (1978). "Modes of perceiving and processing information." *Modes of Perceiving and Processing Information* (H. L. Pick, Jr., and E. Saltzman, Eds.), pp. 1–20. Wiley, New York.

[69] Pick, H. (1987). "Information and effects of early perceptual experience." *Contemporary Topics in Developmental Psychology* (N. Eisenberg, Ed.), pp. 59–76. Wiley, New York.

[70] Stein, B., and Meredith, M. (1993). *The Merging of the Senses*. MIT Press, Cambridge, MA.

[71] Welch, R. B. (1978). *Perceptual Modification: Adapting to Altered Sensory Environments*. Academic Press, New York.

[72] Bower, T. G. R. (1974). "The evolution of sensory systems." *Perception: Essays in Honor of James J. Gibson* (R. B. MacLeod and H. L. Pick, Jr., Eds.), pp. 141–153. Cornell University Press, Ithaca, NY.

[73] Freedman, S. J., and Rekosh, J. H. (1968). "The functional integrity of spatial behavior." *The Neuropsychology of Spatially-Oriented Behavior* (S. J. Freedman, Eds.), pp. 153–162. Dorsey Press, Homewood, IL.

[74] Lackner, J. R. (1981). "Some aspects of sensory-motor control and adaptation in man." *Intersensory Perception and Sensory Integration* (R. D. Walk and H. L. Pick, Eds.), pp. 143–173. Plenum, New York.

[75] Hall, D. L. (1992). *Mathematical Techniques in Multisensor Data Fusion*. Artech House, Boston.

[76] Pavel, M., and Sharma, R. K. (1997). "Model-based sensor fusion for aviation." *Proceedings of SPIE*, **3088**, 169–176.

[77] Hager, G. D. (1990). *Task-Directed Sensor Fusion and Planning: A Computational Approach*. Kluwer Academic, Boston.

[78] Martin, A., Fiscus, J., Fisher, B., Pallet, D., and Przybocki, M. (1997). "System descriptions and performance summary." *Proceedings of the Conversational Speech Recognition Workshop/DARPA Hub-5E Evaluation*. Morgan Kaufman, San Mateo, CA.

[79] Weintraub, M., Taussig, K., Hunicke, K., and Snodgrass, A. (1997). "Effect of speaking style on LVCSR performance." *Proceedings of the Conversational Speech Recognition Workshop/DARPA Hub-5E Evaluation*. Morgan Kaufman, San Mateo, CA.

[80] Oviatt, S. L., MacEachern, M., and Levow, G. (1998). "Predicting hyperarticulate speech during human-computer error resolution." *Speech Communication*, **24**, 87–110.

[81] Banse, R., and Scherer, K. (1996). "Acoustic profiles in vocal emotion expression." *Journal of Personality and Social Psychology*, **70**, 3, 614–636.

[82] Aist, G., Chan, P., Huang, X., Jiang, L., Kennedy, R., Latimer, D., Mostow, J., and Yeung, C. (1998). "How effective is unsupervised data collection for children's speech recognition?" *Proceedings of the International Conference on Spoken Language Processing*, ASSTA Inc., Sydney, Vol. 7, pp. 3171–3174.

[83] Das, S., Nix, D., and Picheny, M. (1998). "Improvements in children's speech recognition performance." *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 433–436. IEEE Press, Menlo Park, CA.

[84] Potamianos, A., Narayanan, S., and Lee, S. (1997). "Automatic speech recognition for children." *European Conference on Speech Communication and Technology*, **5**, 2371–2374.

[85] Wilpon, J. G., and Jacobsen, C. N. (1996). "A study of speech recognition for children and the elderly." *Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP'96)*, pp. 349–352.

[86] Lee, S., Potamianos, A., and Narayanan, S. (1997). "Analysis of children's speech: Duration, pitch and formants." *European Conference on Speech Communication and Technology*, Vol. 1, pp. 473–476.

[87] Yeni-Komshian, G., Kavanaugh, J., and Ferguson, C. (Eds.) (1980). *Child Phonology, Vol. 1: Production*. Academic Press, New York.

[88] Das, S., Bakis, R., Nadas, A., Nahamoo, D., and Picheny, M. (1993). "Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system." *Proceedings of the IEEE International Conference on Acoustic Speech Signal Processing*, Vol. 2, pp. 71–74.

[89] Gong, Y. (1995). "Speech recognition in noisy environments." *Speech Communication*, **16**, 261–291.

[90] Lockwood, P., and Boudy, J. (1992). "Experiments with a non-linear spectral subtractor (NSS), Hidden Markov Models and the projection for robust speech recognition in cars." *Speech Communication*, **11**, 2–3, 215–228.

[91] Junqua, J. C. (1993). "The Lombard reflex and its role on human listeners and automatic speech recognizers." *Journal of the Acoustical Society of America*, **93**, 1, 510–524.

[92] Lombard, E. (1911). "Le signe de l'elevation de la voix." *Annals Maladiers Oreille, Larynx, Nez, Pharynx*, **37**, 101–119.

[93] Hanley, T. D., and Steer, M. D. (1949). "Effect of level of distracting noise upon speaking rate, duration and intensity." *Journal of Speech and Hearing Disorders*, **14**, 363–368.

[94] Schulman, R. (1989). "Articulatory dynamics of loud and normal speech." *Journal of the Acoustical Society of America*, **85**, 295–312.

[95] van Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (1988). "Effects of noise on speech production: Acoustic and perceptual analyses." *Journal of the Acoustical Society of America*, **84**, 917–928.

[96] Potash, L. M. (1972). "A signal detection problem and a possible solution in Japanese quail." *Animal Behavior*, **20**, 192–195.

[97] Sinott, J. M., Stebbins, W. C., and Moody, D. B. (1975). "Regulation of voice amplitude by the monkey." *Journal of the Acoustical Society of America*, **58**, 412–414.

[98] Siegel, G. M., Pick, H. L., Olsen, M. G., and Sawin, L. (1976). "Auditory feedback in the regulation of vocal intensity of preschool children." *Developmental Psychology*, **12**, 255–261.

[99] Pick, H. L., Siegel, G. M., Fox, P. W., Garber, S. R., and Kearney, J. K. (1989). "Inhibiting the Lombard effect." *Journal of the Acoustical Society of America*, **85**, 2, 894–900.

[100] Fuster-Duran, A. (1996). "Perception of conflicting audio-visual speech: An examination across Spanish and German." *Speechreading by Humans and Machines: Models, Systems and Applications* (D. G. Stork and M. E. Hennecke, Eds.), NATO ASI Series, Series F: Computer and Systems Sciences 150, pp. 135–143. Springer-Verlag, Berlin.

[101] Massaro, D. W. (1996). "Bimodal speech perception: A progress report." *Speechreading by Humans and Machines: Models, Systems and Applications* (D. G. Stork and M. E. Hennecke, Eds.), NATO ASI Series, Series F: Computer and Systems Sciences 150, pp. 79–101. Springer-Verlag, Berlin.

[102] Hennecke, M. E., Stork, D. G., and Prasad, K. V. (1996). "Visionary speech: Looking ahead to practical speechreading systems." *Speechreading by*

Humans and Machines: Models, Systems and Applications (D. G. Stork and M. E. Hennecke, Eds.), NATO ASI Series, Series F: Computer and Systems Sciences 150, pp. 331–349. Springer-Verlag, Berlin.

[103] Haton, J. P. (1993). "Automatic recognition in noisy speech." *New Advances and Trends in Speech Recognition and Coding*. NATO Advanced Study Institute.

[104] Senior, A., Neti, C. V., and Maison, B. (1999). "On the use of visual information for improving audio-based speaker recognition." *Proceedings of Auditory-Visual Speech Processing (AVSP)*, 108–111.

[105] Oviatt, S. L. (2000). "Multimodal signal processing in naturalistic noisy environments." *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000)* (B. Yuan, T. Huang, and X. Tang, Eds.), Vol. 2, pp. 696–699. Chinese Friendship Publishers, Beijing.

[106] Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception." *Hearing by Eye: The Psychology of Lip-reading*, (B. Dodd and R. Campbell, Eds.), pp. 3–51. Lawrence Erlbaum, London.

[107] Petajan, E. D. (1987). "An improved automatic lipreading system to enhance speech recognition." Tech. Rep. 11251-871012-111TM, AT&T Bell Labs.

[108] Iverson, P., Bernstein, L., and Auer, E. (1998). "Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition." *Speech Communication*, **26**, 1–2, 45–63.

[109] Oviatt, S. L., Cohen, P. R., and Wang, M. Q. (1994). "Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity." *Speech Communication*, **15**, 3–4, 283–300.

[110] Oviatt, S. L., DeAngeli, A., and Kuhn, K. (1997). "Integration and synchronization of input modes during multimodal human-computer interaction." *Proceedings of the Conference on Human Factors in Computing Systems (CHI'97)*, pp. 415–422. ACM Press, New York.