IMPLICIT USER-ADAPTIVE SYSTEM ENGAGEMENT IN SPEECH, PEN AND MULTIMODAL INTERFACES

Sharon Oviatt
Incaa Designs
821 Second Avenue, Suite 1100
Seattle, WA 98104
oviatt@incaadesigns.org

ABSTRACT

The present research contributes new empirical research, theory, and prototyping toward developing implicit user-adaptive techniques for system engagement based exclusively on speech amplitude and pen pressure. The results reveal that people will spontaneously adapt their communicative energy level reliably, substantially, and in different modalities to designate and repair an intended interlocutor in a computer-mediated group setting. Furthermore, this sole behavior can be harnessed to achieve system engagement accuracies in the 75-86% range. In short, there was a high level of correct system engagement based exclusively on implicit cues in users' energy level during communication.

KEY WORDS

Adaptive systems, handwriting recognition, pressure effects, speech recognition, user interface human factors.

INTRODUCTION

Recent interface design has placed increasing emphasis on developing mobile, educational, and other applications that minimize cognitive load on users, so they can remain focused on demanding field tasks. The present research contributes new empirical research, theory, and prototyping toward developing implicit user-adaptive techniques for system engagement based exclusively on speech amplitude and pen pressure. First, empirical information was collected on users' spontaneous increase in energy (i.e., vocal amplitude, manual pressure) when communicating using the speech and pen modalities to engage a computer versus human partner during computer-mediated collaborative meetings. Based on preliminary analyses of changes in users' communicative energy during spoken and written interaction, a formula was developed for deriving user-specific thresholds to automatically distinguish computer- from human-directed input in real time while users spoke or wrote throughout actual meetings. This research reports on users' spontaneous changes in communicative energy, as well as further adaptations in their energy over time after interacting with a simulated implicit engagement system that provided error feedback contingent on whether their energy was above or below a habitual threshold level. It also summarizes the impact of this simulation on system engagement reliabilities for the speech amplitude and pen pressure techniques. Finally, it investigates the extent to which users were aware of changes in their own energy level when using this kind of implicit interface, as well as the impact of engaging a tutorial system over 100 times per session on their ability to solve complex mathematics problems correctly. In this respect, the study investigated whether this type of interface could be successfully implemented, while remaining transparent to users and avoiding distracting them and jeopardizing performance.

Research Strategy & Philosophy & Challenges

The main elements of the present research approach were to (1) model and accommodate users' natural communication patterns, because many aspects are highly engrained and not under full conscious control (e.g., timing, amplitude), so they would be difficult or impossible for people to unlearn. As such, interfaces incompatible with their natural behavior would precipitate more system errors and be less usable; (2) leverage users' subconscious and over-learned behavior patterns to minimize cognitive load and enhance performance; (3) provide users with functionality that they are strongly motivated to achieve, in this case being recognized correctly by an intended interlocutor; (4) design user-adaptive interfaces tailored to individual users so system reliability can be optimized, especially for communication technologies since users' communication patterns are subject to large individual differences.

Related Theoretical Work

Lindblom et al. have formulated the H & H theory to account for stylistic variation in interpersonal speech. This theory asserts that speech signal adaptation varies actively along a continuum from *hypo- to hyper-clear speech* [2].

Hypo-clear speech is relatively relaxed, and involves minimal expenditure of articulatory effort by the speaker, instead relying on the listener's ability to fill in missing signal information from knowledge. In contrast, hyper-clear speech is a clarified style based on greater energy expenditure, so it is more intelligible and relies less on knowledge. Manifestations of increased vocal effort include production of ideal target values for the acoustic form of vowels and consonants, and higher amplitude speech.

Lindblom and colleagues have argued that speakers make a moment-by-moment assessment of their listener's need for explicit signal information, and they adapt their speech production to the perceived needs of a particular listener in

context. Essentially, Lindblom believes that speakers operate on the *principle of supplying sufficient discriminatory information* for a listener to comprehend their intended meaning, while at the same time striving for articulatory economy. Hypo-clear speech is the default speaking style, but when a threat to comprehension is anticipated or actually experienced (e.g., noisy environment) then the speaker will adapt to hyper-clear speech. Speakers also routinely engage in hyperarticulate speech with computers, because they expect them to be error-prone [5]

Apart from these speech adaptations that enhance the intelligibility of semantic content, many animals and humans also increase amplitude to call distant group members, and to attract and maintain the attention of nearby interlocutors. These changes in speech amplitude trigger involuntary attentional shifts in the brain of listeners, supporting their ability to orient to and correctly identify an intended interlocutor so lexical content can be processed successfully [6]. Recent empirical research indicates that a user's amplitude level is a strong marker of whether she is talking to herself, a peer, or a computer in a computer-mediated meeting, with substantial progressive amplitude increases in each of these cases [3]. Although there are different amplitude ranges for addressing different types of interlocutor, people have limited awareness of these dynamic changes in their own amplitude as they speak.

Lindblom's H&H theory accounts for dynamic speech signal adaptations that fortify lexical meaning in interpersonal contexts, especially articulatory changes. The present research builds from this theoretical framework by asserting that adaptations in communicative effort along the hypo-to-hyper spectrum are characteristic of *all modes of communication.* That is, they are not modality specific. For example, it is conjectured that writers also will expend more effort to clarify their input when they believe its intelligibility is threatened, including increasing their pressure when interacting with computers. The present work also generalizes Lindblom's theory to include *interactive computer exchanges*, not just interpersonal ones. Finally, it generalizes the applicability of this theory more broadly than simply conveying lexical meaning to other acts such as *designating an intended interlocutor.*

Study Goals

The primary objectives of the present study involved empirical research and prototyping of implicit user-adaptive interfaces involving speech and pen input for collaborative use in field settings. The theoretical framework upon which this study is based, its simulation methodology and research strategy, and its empirical findings all constitute unique developments in human interface research. The following questions were examined:

- Do people spontaneously adapt the energy level of their communications to distinguish addressing a computer versus human partner during computer-mediated group meetings? If so, is this manifest as *higher amplitude levels* when addressing the computer during speech interactions, and *higher pressure levels* during pen-based interactions?
- Can a user-adaptive system be designed for speech and/or pen input that yields reliable system engagement *entirely implicitly* based on these naturally-occurring energy differences?
- If a system adapted to users' natural communication patterns is deployed, then will they recognize these system response contingencies and *further adapt their behavior to optimize system reliability*? If so, will this further adaptation be manifest as (1) increased energy when addressing the computer, (2) decreased energy when addressing a human, or both? And will such user adaptation lead to (3) greater differentiation in their energy over time when addressing a computer versus human, or (4) higher system reliabilities by the end of a training session?
- If people use increased energy to mark a computer addressee, then will people forcefully increase their amplitude or pressure as a repair strategy following a system failure to recognize that it was addressed?
- Can this type of reliable system engagement occur in spite of limited awareness by users' of their behavior?
- Does this type of implicit system engagement avoid distracting users, such that cognitive load remains low and performance is preserved during demanding tasks?

METHODS

Subjects, Tasks & Procedure

In this research, data were collected with 12 pairs of students who solved complex math problems using a tutorial system that they engaged over 100 times per session *entirely implicitly* via speech amplitude or pen pressure cues. Each session consisted of 15 math problems presented as word problems.

Each pair of students participated in two sessions, one involving spoken interaction with the simulated tutoring system, and the other written interaction. Students had a calculator, close-talking microphones for speech input, and digital pens with multiple sheets of digital paper for pen input and working out the problems. During each of the sessions, participants were instructed to solve 4 practice problems, and 15 problems during the main session. Both participants were encouraged to use the digital paper and pen to write "scratch" notes while working on problems. They were told to discuss each problem together and be sure they understood it and could explain their answers, since they would be asked to do so. One student was designated to interact with the computer whenever assistance was needed. This student was the one who gave spoken or written requests to the computer.

During practice, students became familiar with the type of problems and tutoring system interface. This phase also permitted collection of baseline data to establish the students' amplitude threshold during speech sessions and their pressure threshold during pen sessions. At the end of each session, the student who had interacted with the tutoring system was interviewed about the system, its errors, their awareness of changing their speech amplitude and pen pressure when addressing the computer. Each session took about 1.5 hours to complete.

Dual-Wizard Simulation Environment

To simulate a system that could be automatically engaged entirely through *implicit* user communication cues, with no "moding" or explicit instruction of any kind from the user, a dual-wizard method was implemented. With this environment, the wizard could view multiple video feeds of the group's interaction as data was collected. Each participant's writing was collected using a Logitech digital pen and multiple large sheets of Anoto paper, and digital ink was streamed live to a virtual canvas which could be panned, rotated, and zoomed while the wizards responded. Synchronized and time-stamped data also was collected of each student's speech and written input as they worked using Countryman close-talking hyper-cardioid microphones connected to Shure wireless transmitters and receivers.

Based on the semantic content of a student's speech or pen input, the first wizard's role was to identify whether a construction was intended as a request to the computer. This judgment was based on the presence of key phrases, words, or diagrams (e.g., "Rhombus" or a diagram of one would prompt a definition of that term). During student conversations, a key phrase or diagram could be associated with a computer-directed utterance, but sometimes they could occur spuriously during interpersonal discussion. The second wizard's role was to track signal features (e.g., speech amplitude, pen pressure) of constructions flagged by the first wizard as potentially computer-directed to determine whether they also met a *user-defined threshold* required for responding to them as computer-directed. In summary, during students' conversation about their math problems their utterances were: (1) *filtered for semantic relevance* to the tutoring system's application functionality, and then (2) *filtered for communicative energy (amplitude, pressur*e), so a decision could be made about whether the system should acknowledge a particular construction as computer-directed or not. For details of this novel automated dual-wizard environment, its functionality, implementation, response capabilities, and visual interface see [1]. For details of calculating user-centered thresholds, see [6].

Error Generation & Contingent Responding

Whenever a user construction met the semantic and energy criteria for requesting computer functionality, the wizards responded to that "target" utterance as computer-directed and the user received a correct computer response, resulting in a HIT. However, if amplitude was below threshold, then no response was delivered and the user's request was ignored, resulting in a MISS. In this case, the computer responded with "I'm sorry, I didn't catch that," and users repeated their request. In other cases, an utterance could meet semantic criteria although it was intended for a human peer. If the user's amplitude threshold was not exceeded, the wizard ignored it, resulting in a CORRECT REJECT. However, if the threshold was exceeded, then the wizard responded as if the utterance was computer-directed, in which case a FALSE ALARM was produced and the computer intruded with "What can I do?" In summary, the simulated system responded as a real implicit user-adaptive system would with respect to error pattern. This provided an opportunity for users to learn from the error pattern during contingent system responding by further differentiating their energy.

Research Design

The main within-subject independent factors included: (1) Modality of interaction (Speech, Pen), and (2) Intended addressee (Computer, Human). Half of student pairs completed their speech session first, and the other half written input first. Since adaptations in users' communication patterns and system responding were evaluated over the session, problems also were presented in both forward and reverse orders (i.e., 1-15, versus 15-1), with half of participants in each condition receiving each presentation order.

RESULTS

The analyses reported here were based on all spoken data throughout the speech session (approximately 1600 spoken utterances) and written data in the pen session (1400 written constructions) for the student designated to interact with the system, as well as self-report data following those sessions. Analyses of adaptations following system misses were based on approximately 130 matched pairs of spoken utterances and 270 matched pairs of written constructions.

Speech Energy: Amplitude Findings

Figure 1 shows the average speaker amplitude for all 12 pairs when speaking to the computer versus a human partner from the baseline period through problem triad 5, as well as the average user-centered threshold level. During the baseline period before any user-centered amplitude threshold was applied, speakers' spontaneous average amplitude when addressing the computer was 62.45 dB, significantly higher than 56.97 dB when addressing a human peer, paired t (6) = 7.26, $p < .001$, one-tailed. After the amplitude contingency became active, speakers' average amplitude when addressing the computer increased from 62.45 during baseline to 63.20 on triad 5 at the end of the session, a marginally significant increase by paired t test, t (11) = 1.53, $p < .077$, one-tailed. Speakers' average amplitude decreased when addressing their human collaborator from 58.97 to 57.44 at the end of the session, a significant decrease by paired t test, t (9) = 2.21, $p < .027$, one-tailed. As a result, there also was a significant expansion of the differential in amplitude between computer- and human-directed speech from 4.44 dB on triads 1 and 2 to 5.75 dB on triads 4 and 5 at the end, paired t (9) = 1.80, $p < .052$, one-tailed.

During error handling, speakers increased their average amplitude from 59.4 dB immediately before a miss to 62.7 dB afterwards, a significant increase by paired t (7) = 9.66, $p < .001$, one-tailed. This 3.31 dB difference represented a 46.4% increase in linear energy following a computer miss. Furthermore, 100% of students increased their amplitude when resolving misses.

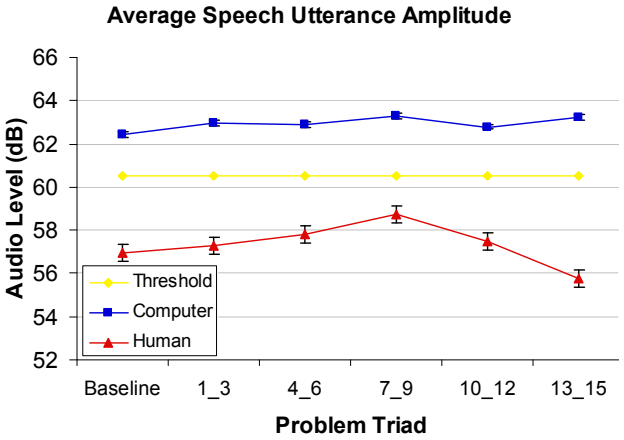

Figure 1: Average amplitude in dB over the session for computer- versus human-directed spoken utterances

Writing Energy: Pressure Findings

Figure 2 shows the average pressure for all 12 pairs when writing to the computer versus a human partner from baseline through triad 5, as well as the average pressure threshold level. During the baseline period, writers' spontaneous average pressure when addressing the computer was .947, significantly higher than .923 when addressing a human peer, paired t (11) = 3.58, $p < .002$, one-tailed. After the pressure contingency became active, writers' average pressure when addressing the computer increased from .947 during baseline to .952 on triad 5, a significant increase by paired t test, t (11) = 2.95, $p < .007$,one-tailed. However, neither writers' average pressure when addressing their human collaborator nor their pressure differential between computer- and human-addressed input changed significantly across the session, paired t (9) < 1.

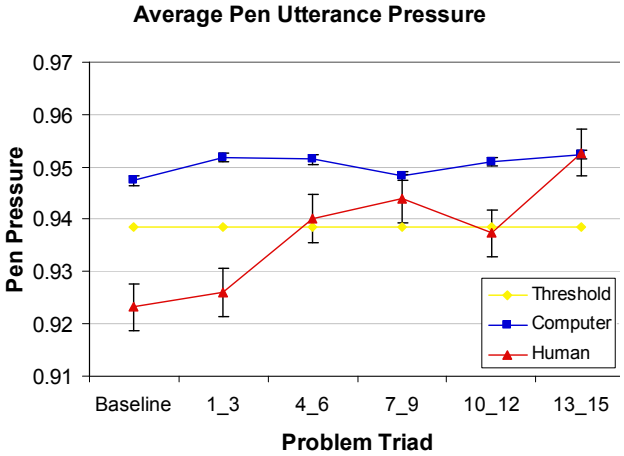

Figure 2: Average pen pressure over the session for computer- versus human-directed written constructions

During error handling, writers increased their average pressure from .923 before a miss to .943 afterwards, a significant increase, paired t (7) = 4.93, $p < .001$, one-tailed. This .021 difference represented a 9.5% increase in energy.[2] Once again, 100% of students increased their pen pressure when resolving these computer misses.

System Reliability

For all 24 sessions, the average reliability of correctly engaging the system based on speech amplitude and pen pressure was well above 50% chance level overall. In the speech sessions, 7 of 12 students achieved reliabilities in the 90-100% range, and 11 of 12 in the 70-100% range. Using pen pressure, 10 of 12 subjects had reliabilities in the 70-100% range. The average system reliability achieved by triad 5 at the end of the speech sessions was 86.0%, whereas for the pen sessions it was 75.2%, a significant difference by paired t test, t (11) = 2.09, $p < .031$, one-tailed.

During speech sessions, average system reliability improved from 82.6% on problem triad 1 to 86.0% on triad 5, or 3.4%. This improvement represented a 24.3% relative reduction in the speech error rate from the beginning to end of the session, which primarily was due to reduction in false alarms as speakers dropped their amplitude to their human partner. However, during pen sessions average system reliability did not show improvement during this 1-hour interval.

Self-Report on Communicative Energy

For speech, only 41.7% of people spontaneously mentioned talking louder to the computer during three open-ended interview questions on this topic after their session, whereas 50.0% acknowledged talking louder to the computer when specifically asked whether they did so. For pen input, 0% spontaneously mentioned writing more forcefully or with greater pressure to the computer, and just 8.3% acknowledged doing so when specifically asked. A comparison of positive responses when prompted confirmed greater user awareness of their speech amplitude changes than pen pressure, $\chi^2(1) = 9.25$, $p < .01$.

Maintenance of Performance Level

When using an interface involving implicit pen pressure to engage the system, students' correct problem solutions averaged 66.83% during the first seven problems and 72.33% during the last seven, not a significant change in correct solutions, $t < 1$, N.S. When using an interface based on speech amplitude to engage the tutoring system, correct problem solutions averaged 78.00% during the first seven problems and 79.75% on the last seven problems, again not a significant change, $t < 1$, N.S. As such, no deterioration in students' performance was observed across any sessions.

Overall, students' problem solutions averaged 79.46% correct during system engagement using speech amplitude, but only 70.24% during the pen pressure engagement, which was a significantly higher performance level on solving math problems using the speech engagement method, paired t (11) = 2.58, $p < .026$, two-tailed.

DISCUSSION

In summary, these results reveal that people will spontaneously adapt their communicative energy level reliably, substantially, and in different modalities to designate and repair an intended interlocutor in a computer-mediated group setting. Furthermore, this sole behavior can be harnessed to achieve system engagement accuracies in the 75-86% range, which would be especially valuable for mobile communication technologies. Overall, 86% of the time (i.e., 6 times out of 7) there was correct engagement of the computer based exclusively on implicit changes in their speech amplitude. Likewise, 75% of the time (i.e., 3 times out of 4) the computer was correctly engaged based exclusively on implicit changes in users' manual pressure when writing. In short, there was a high level of correct system engagement based exclusively on implicit cues in users' energy level during communication.

Although students used these interfaces to engage a tutoring system over 100 times during their sessions, they nonetheless reported limited or no awareness of using amplitude or pressure to control the interface. Based on spontaneous self-reports gathered after their sessions, no students mentioned using greater pen pressure when providing input or correcting errors with the computer, and less than 42% mentioned using greater volume when speaking. Furthermore, on these complex mathematics problem solving tasks, students were able to maintain their performance level without deterioration throughout a lengthy session. However, the interface operated via speech amplitude, which had the substantially lower 14% error rate, supported an average of +9.22% higher correct problem solutions than the pen pressure interface. In summary, effective interfaces can be designed based on implicit cues that do not require users' awareness or focused attention at all, so that distraction from their primary task can be minimized.

From a theoretical standpoint, this research substantially generalizes Lindblom's theory by asserting that adaptations in communicative effort along the hypo-to-hyper spectrum are characteristic of *all modes of communication*, not simply speech. These adaptations also

[2] Linear speech energy was calculated using the transformation $A' = .00002*10^{A/20}$ – where A is the speech amplitude in dB and A' is the linear speech energy. Linear pen pressure was calculated using the transformation $P' = .227\ln(P) + .9367$ – where P is the pen pressure value from the Anoto™ pen and P' is the linear pen pressure (force) in Newtons.

are characteristic of *human-computer communications*, not just interpersonal ones. Finally, they extend beyond conveying lexical meaning to communicative acts like *designating an intended interlocutor.*

As more emphasis is placed on developing mobile, educational, and other applications that exert minimal cognitive load on users, it will become essential to explore interfaces based on implicit engagement so users can remain focused on their primary field tasks.

REFERENCES

1. Arthur, A., Swindells, C., Oviatt, S. and Cohen, P. A high-Performance dual-wizard infrastructure supporting speech and digital pen input, in submission.
2. Lindblom, B. Explaining phonetic variation: A sketch of the H and H theory, *Speech Production and Speech Modeling* , ed. by W. Hardcastle and A. Marchal, Kluwer, Dordrecht (1990), 403–439.
3. Lunsford, R., Oviatt, S. and Arthur, A. Toward open-microphone engagement for multiparty interactions, *Proc. of the International Conference on Multimodal Interfaces,* ACM Press (2006), 273-280.
4. Messer, D. The identification of names in maternal speech to infants. *Journal of Psycholinguistic Research*, *10* (1), (1981), 69-77.
5. Oviatt, S., MacEachern, M. and Levow, G. Predicting hyperarticulate speech during human-computer error resolution, *Speech Communication* (1998), 24 (2), 1-23.
6. Oviatt, S., Swindells, C. and Arthur, A. Implicit user-adaptive system engagement in speech and pen interfaces, in submission.
7. Schroger, E., A neural mechanism for involuntary attention shifts to changes in auditory stimulation. *Journ. of Cognitive Neuroscience*, *8*(6), 1996, 527-539.