Implicit User-Adaptive System Engagement in Speech and Pen Interfaces

Sharon Oviatt & Colin Swindells
Incaa Designs
821 Second Ave., Ste. 1100, Seattle WA. 98104
oviatt@incaadesigns.org

Alex Arthur
Adapx
821 Second Ave., Ste. 1150, Seattle WA. 98104
alex.arthur@adapx.com

ABSTRACT

As emphasis is placed on developing mobile, educational, and other applications that minimize cognitive load on users, it is becoming more essential to base interface design on implicit engagement techniques so users can remain focused on their tasks. 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. Results revealed that users spontaneously, reliably, and substantially adapted these forms of communicative energy to designate and repair an intended interlocutor in a computer-mediated group setting. This behavior was harnessed to achieve system engagement accuracies of 75-86%, with accuracies highest using speech amplitude. However, students had limited awareness of their own adaptations. Finally, while continually using these implicit engagement techniques, students maintained their performance level at solving complex mathematics problems throughout a one-hour session.

Keywords
System engagement, implicit, user-adaptive, speech amplitude, pen pressure, educational interface

ACM Classification Keywords
H5.2 Information interfaces and presentation: User interfaces— user-centered design, theory and methods, interaction styles, input devices and strategies, evaluation/methodology, voice I/O, natural language.

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 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 the 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. Perhaps the greatest challenge of pursuing this research strategy

involves modeling users' natural communication patterns, including variation in speech amplitude and pen pressure which is extremely data-intensive. Furthermore, correctly distilling the main parameters that drive these sources of variation, and then applying this information to interface design, is an unsolved problem.

Related Theoretical and Empirical 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* [6] 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 [6]. 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. During individual human-computer interaction, this typically is manifest as durational and articulatory effects, rather than amplitude changes [12].

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 [9]. The cognitive neuroscience literature confirms that these changes in speech amplitude trigger involuntary attentional shifts in the brain of listeners [1], supporting their ability to orient to and correctly identify an intended interlocutor so lexical content can be processed successfully. 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 [8], with substantial progressive amplitude increases in each of these cases [7]. 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.*

Additional Related Work on Interfaces for Pen Moding & Speech Open-Microphone Engagement

The present research is related to past work on "open-microphone engagement." Rather than requiring explicit keywords or push-to-talk, which is widely known to be error-prone and a usability problem, open-microphone engagement techniques have been based on semantic and audio-visual information sources such as speaker head position and gaze direction, and the presence of lip movements and articulated speech [10, 13, 18]. In spite of applying interesting engineering techniques (e.g., stream-weighting estimation, Bayesian decision theory, multilayer perceptrons), one problem with existing open-microphone engagement is that the critical information sources still have not been adequately modeled. For example, users frequently direct their gaze at a computer when actually talking to a human peer. Users also often face a computer while talking with their lips moving in cases where they are self-talking [7]. Another issue is that open-microphone engagement has mainly been explored for individual human-computer interaction, rather than for computer-mediated meetings in actual field settings— which presents the far more difficult problem addressed in this study.

This research also is related to the literature on pen interface "modes," a classic interface problem which recently has been inspired by interest in improving pen tablet interfaces [3, 5, 14, 15]. This literature has mainly been focused on developing mechanical and sometimes two-handed engagement techniques for large gesture sets that include inking, erasing, selection, highlighting, object creation, etc. [3], rather than simply distinguishing when writing is intended as private note-taking versus a command to the computer, which is the most common basic distinction required for digital paper and pen interface design. The design of "modeless" pen interfaces has included promising inferential methods, although they involve interruptive prompting [15]. It also has included pressure-based engagement techniques that could be deployed with as many as 6 distinct levels [14], and which already have been implemented to control stroke thickness in drawing programs. However, pressure engagement

techniques have not yet incorporated user-centered thresholds, which clearly could improve their overall usability [5]. In short, there still has been little systematic exploration of the cognitive science basis or potential usability of pressure-sensitive pens for different purposes (although see Kao et al. [4]), especially in the emerging area of digital paper and pen interface design.

Study Goals and Hypotheses

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 specific questions and related hypotheses were examined:

1. 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?
Hypothesis: People will use both higher speech amplitude and higher pen pressure levels to mark a computer rather than human as an intended addressee.
2. 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 so, what level of reliability can be achieved?
Hypothesis: Reliable system engagement can be achieved above chance levels based on implicit energy cues in users' communication, with speech amplitude yielding higher reliabilities than pen pressure.
3. 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?
Hypothesis: Over a session, people will increase their amplitude and pressure when addressing the computer, decrease them when addressing a human, and expand their energy differential between computer and human—which will yield improvement in overall system reliability.
4. 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?
Hypothesis: Following computer failures to engage, people will forcefully increase both speech amplitude and pen pressure as an adaptation to repair the error. Basically, they will accentuate their naturally-occurring behavior.

5. Can this type of reliable system engagement occur in spite of limited awareness by users' of their own behavior?
Hypothesis: Most users will be unaware of changes in their own energy level when addressing a computer or repairing errors. They will be more aware of amplitude than pressure changes, since speech is publicly observable.
6. Does this type of implicit system engagement avoid distracting users, such that cognitive load remains low and performance is preserved on hard problem solving tasks?
Hypothesis: The percent of correct math problem solutions will remain the same or even improve over the session, because the cognitive load associated with an implicit engagement technique is minimal. If one technique has a higher reliability level than the other, then it will be associated with lower load and higher performance levels.

METHODS

Subjects

Twelve high school student groups participated in the study, with two students matched on gender and mathematics ability constituting each group. Seven groups were female and 5 male, and 9 were high performers versus 3 low performers in geometry. All students were 14 to 18 years old, and had recently completed Geometry 1 at a Seattle-area high school. Participants were paid volunteers, free of disabilities, and all were native English speakers.

Tasks

Each session consisted of 15 basic algebra and geometry problems presented as word problems. Figure 1 shows an example geometry word problem and its solution. To ensure the generality of results, two parallel problem sets were developed. Each set included five groups of three problems apiece involving low, moderate, and high difficulty. The difficulty levels were validated using teacher records of percent correct on these problems for high school students in introductory geometry, and also pilot data.

Daniel is building a half pipe so he can do bike tricks. If he wants the radius of the pipe to be 16 feet, how long should the plywood be for the curved inner area?

Answer: 50.27 feet (16π)

Example Solution:
  Circumference of a circle = 2π (radius)
  Circumference of 1/2 circle = π (radius)

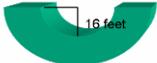Circumference = π (16) feet
              = 50.27 feet

Figure 1: Example geometry problem and solution

Procedure

Each pair of students participated in two sessions, one involving spoken interaction with the simulated tutoring system, and the other written interaction. For each session

in this longitudinal sequence, the two students were seated at a table together. They had a calculator, close-talking microphones for speech input, and digital pens with multiple sheets of A1 digital paper (594 mm x 841 mm) for pen input and working out the problems. Digital pen input was color-coded to identify each student's contributions.

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, and they were asked to print their input. 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. After giving instructions and initial practice, the experimenter left the room and students were told they could access the experimenter for assistance if needed.

At the end of each session, the student who had interacted with the tutoring system was interviewed about the system and any errors it made. They were asked about the type of errors, their frequency, and whether they changed over the session. They were also asked why they thought errors occurred during speech and pen input, and whether there was anything they did that helped avoid or recover from errors. Initial questions were open ended, but became more specific in asking whether students thought they were using speech amplitude or pen pressure to reliably engage the computer. Each session took about 1.5 hours to complete.

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.

Computer-Assisted Instruction Application
A computer tutorial assistant running on a desktop PC was placed on the table in front of students as they worked, and students were told they could use it to assist them as they solved problems together. This system was used to display each math problem, provide information about relevant terms and equations, submit their answers, request information about the correct answers, and also explanations of problem solutions. Information was provided in visual form on the monitor, sometimes with text-to-speech (e.g., explaining definitions of math terms or equations). A summary of available computer functions was provided on the bottom of the computer monitor for

students' reference. They were told that they could use either speech or pen input to interact with the tutorial system, and during one session each group used speech input while in the other they used pen input.

Dual-Wizard Environment
The computer assistant was implemented using a dual-wizard simulation, which was presented to participants as a fully functional system. Simulation infrastructure was developed for this study that permitted the wizard to view multiple video feeds of the group's interaction as data was collected, including a close-up view of each student, a wide-angle view of the room in which they were working, and a close-up of the tabletop with a view of each student's written input while they solved math problems. 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. Countryman close-talking hyper-cardioid microphones, which students were instructed not to touch, were connected to Shure wireless transmitters/receivers to collect speech input from individual students. In addition, a digital audio recording of group conversation was collected using a studio-quality omni-directional microphone hung above the table at which students worked. Finally, each participant's writing was collected using a Logitech digital pen and multiple large A1 sheets of Anoto paper. Details about the data capture and synchronization of the audio-visual-pen media streams is presented elsewhere [1, 2].

During the simulation, the wizards could listen to either the student team leader's speech, or a general room microphone of group discussion. The wizards also could zoom in on a given student's written input on a particular sheet of paper to see it more closely, and could rotate its orientation for clear visibility in tracking what students were writing. 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. During speech input, software: (1) automatically segmented a speaker's real-time speech into individual utterances, (2) displayed an utterance's average amplitude, and the change in average amplitude between it and the preceding utterance, and (3) provided a color-coded (red/green) indication to the wizard of whether an utterance was above

or below that speaker's amplitude threshold. During pen input, the wizard's role was analogous except that: (1) a pen tablet computer was used by the wizard to manually encircle individual pen constructions that met semantic criteria. The simulation software then automatically (2) displayed that construction's average pressure, and (3) provided an indication of whether it met that writer's pressure threshold. Based on this streaming real-time information for speech and pen input, the wizard could simulate a system that either responded to the speaker's request (i.e., "hit") or ignored it (i.e., "miss"), contingent upon semantic content and energy level.

In summary, using this dual-wizard technique, during students' conversation about their math problems utterances were: (1) *filtered for semantic relevance* to the tutoring system's application functionality, and then (2) *filtered for communicative energy*, so a decision could be made about whether the system should acknowledge the construction as computer-directed or not. The simulation environment supporting these wizard roles and their coordination was designed to provide very speedy and accurate responding. For example, the average dual-wizard response time was .98 seconds during speech interaction, and 1.63 seconds for pen. For details of this novel simulation environment, its functionality, implementation, response capabilities, and visual interface, see [2].

Calculation of User-Centered Thresholds

During the practice phase, a given student's speech or pen input was analyzed on an utterance-by-utterance basis without any training. Average amplitude or pressure was computed to establish that student's user-centered range, average values when addressing a human peer versus the computer, and *user-centered threshold* for distinguishing when that user was addressing the computer such that the system should automatically engage and respond. Each utterance was classified as computer-directed if it was above their threshold, or as human-directed if below.

Speech Amplitude Threshold Calculation: The amplitude threshold during speech input could be met by reaching *either the average amplitude or the differential amplitude criteria*. Equation 1 is the formula for calculating a student's *average amplitude*, A_{Amp}. Equation 2 is the formula for calculating *average amplitude differential* from previous to present utterance, D_Δ.

The above decision criteria were developed based upon empirical analysis of the "acoustic scene" typical of meetings involving teenagers solving math problems. These criteria needed to suppress false alarms in open-microphone engagement which otherwise would be caused by amplitude spikes during overlapped speech and laughter. Over 30% of all utterances involved overlapped speech, and over 10% laughter. These sources of acoustic "noise" are unique to group exchanges, and similarly high rates of overlapped speech have been reported during adult meetings [17].

$$A_{Amp} = \bar{x}_{Amp} + 0.5\sigma_{Amp} \tag{1}$$

A_{Amp} = Amplitudes above which utterance is assumed computer-directed, below human-directed. $\bar{x}_{Amp}$, σ_{Amp} = Mean and standard deviation of amplitude on all computer-directed, human-directed and other utterances.

$$D_\Delta = \bar{x}_\Delta + 0.5\sigma_\Delta \tag{2}$$

D_Δ = Amplitude difference between two consecutive utterances, with values above assumed computer-directed, below human-directed. $\bar{x}_\Delta$, σ_Δ = Mean & standard deviation of utterance amplitude differentials, D_Δ, for all computer-directed, human-directed and other utterances.

Pen Pressure Threshold Calculation: Equation 3 is the formula for calculating a student's pen pressure threshold while writing, T_{Pen}. During real-time processing, the raw pressure data for user constructions were filtered by removing the first and last two data points from each stroke to eliminate spikes due to initial tapping down on the writing surface. Pen strokes with fewer than 10 data points also were filtered out. Pressure averages then were calculated based on the first 30 points sampled during a construction, due to better user control of pressure then and also better separation between users' computer and human pressure. Decision criteria for pressure thresholds were based on empirical analysis of pressure during pilot testing.

$$T_{Pen} = \bar{x}_{Cmpt} - 0.5\sigma_{Cmpt} \tag{3}$$

T_{Pen} = Pressure above which construction is assumed computer-directed, below human-directed. $\bar{x}_{Cmpt}$, σ_{Cmpt} = Mean and standard deviation of pressure on all computer-directed pen utterances.

Error Generation Capabilities & Contingent Responding

As discussed earlier, 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, after practice was over 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 errors during

contingent system responding by further differentiating their energy.

Research Design

The order of speech and pen sessions was counterbalanced, with half of student pairs completing their speech session first and the other half written input. Half of the speech and pen sessions received problem set one, and the other half set two. 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.

The main within-subject independent factors included: (1) Modality of interaction (Speech, Pen), and (2) Intended addressee (Computer, Human). The two problem sets (1, 2) and their difficulty level (low, moderate, high) were included to ensure generality and realism of the results, as was sampling of both high and low student performers. Inclusion of counterbalancing for modality presentation order and problem set presentation order primarily was to ensure experimental precision and control.

Data Analysis and Dependent Measures

Time-stamped and synchronized data were collected on students' speech and written input throughout both sessions. 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.

Speech Amplitude & Pen Pressure
Average speech amplitude when addressing the computer and human with semantically-relevant utterances was summarized for triads 1-5 for all speech sessions, as was the average differential in amplitude. Average pen pressure when addressing the computer and human with semantically-relevant constructions also was summarized for triads 1-5 for all pen sessions. Matched pairs of speech utterances before and after system misses were analyzed for change in amplitude, and similar matched pairs of pen constructions were analyzed for change in pressure.

System Reliability
Average system reliability was summarized for all triads for speech and pen sessions, as were misses and false alarms.

Self-Report on Communicative Energy and Errors
The percent of subjects reporting awareness that they used greater speech amplitude or pen pressure when addressing the computer was summarized, including both spontaneous self-reports and when asked explicitly. The percent of subjects who correctly reported change in the system error rate during speech and pen sessions also was summarized.
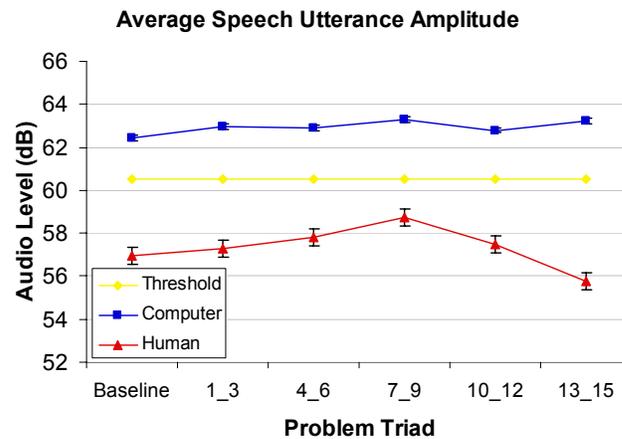

Average Speech Utterance Amplitude

Figure 2: Average amplitude in dB over the session for computer- versus human-directed spoken utterances

Maintenance of Performance Level
The percent of correct problem solutions was summarized for problem triads 1-5 for all speech and pen sessions.

RESULTS

Speech Energy: Amplitude Findings

Figure 2 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. In comparing amplitude changes when addressing a human, problem triads 1 and 2 and also 4 and 5 were collapsed, due to more limited data on semantically filtered human-directed speech. 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.

Writing Energy: Pressure Findings

Figure 3 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
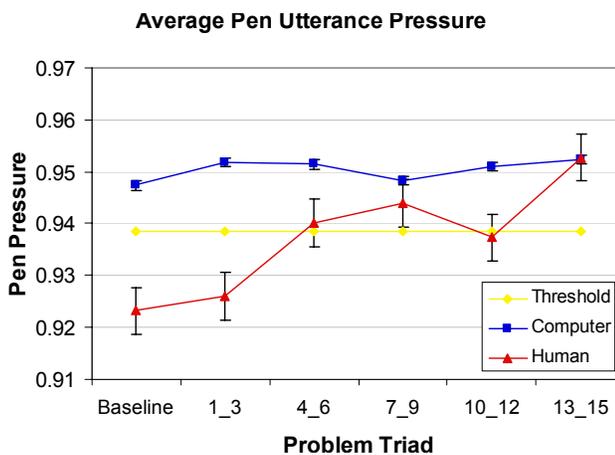
Average Pen Utterance Pressure

Figure 3: Average pen pressure over the session for computer- versus human-directed written constructions

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. Writers' average pressure when addressing their human collaborator was .937 during problem triads 1 and 2 and .944 on triads 4 and 5 by the end, not a significant difference by paired t test, t (9) < 1. Further analysis also revealed no significant change in the pressure differential between computer- and human-addressed input across the sessions, paired t (9) < 1.

System Reliability
For all 24 sessions, the average reliability of correctly engaging the system based on amplitude and pressure was well above 50% chance level overall. Individual differences in reliabilities by triad 5 for the speech session ranged from 62.1%-100% for the 12 pairs, and reliabilities for the pen sessions ranged from 54.5%-87.5%, as shown in Figures 4 and 5. 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 changed from 78.6% on triad 1 to 75.2% on triad 5, or -3.4%, indicating no improvement in this 1-hr. period.
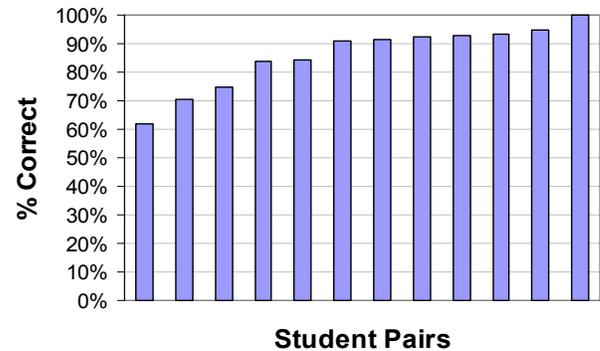
Triad 5 System Reliability

Figure 4: Average system reliability on triad 5 based on simulation of user-centered speech amplitude thresholds
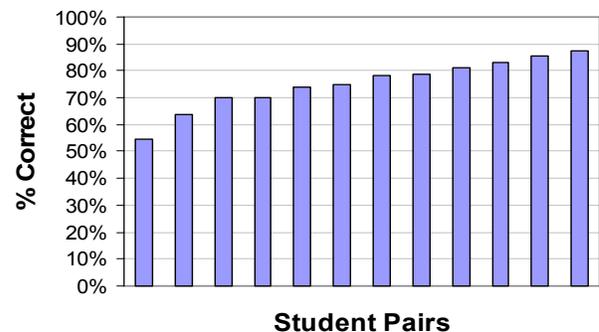
Triad 5 System Reliability

Figure 5: Average system reliability on triad 5 based on simulation of user-centered pen pressure thresholds

Energy Changes during Error Handling
Approximately 130 matched pairs of spoken utterances and 270 matched pairs of written constructions were available for analysis immediately before and after system "misses" throughout the main session, based on 8 of 12 pairs for whom at least 5 misses were available. During speech sessions, 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. As observed in Table 1, 100% of students increased their average amplitude following a miss.

Likewise, during the pen sessions 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] As observed in Table 1, 100% of students increased their average pressure following a miss.

---

[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

Table 1: Increased energy before and after computer misses

Pair	Speech Amplitude (dB)				Pen Pressure			
	Pre	Post	Diff	% Energy Increase	Pre	Post	Diff	% Energy Increase
1	57.3	59.1	1.8	22.5%	.922	.929	.007	3.2%
2	54.8	57.7	2.9	39.3%	.920	.929	.009	4.0%
3	60.6	63.6	2.9	40.3%	.923	.934	.011	4.9%
4	60.9	64.0	3.1	42.8%	.923	.938	.014	6.5%
5	59.8	63.0	3.2	45.2%	.924	.950	.026	12.2%
6	59.6	63.0	3.4	48.2%	.921	.948	.027	12.4%
7	60.8	64.9	4.1	59.8%	.922	.953	.032	15.0%
8	61.4	66.5	5.1	80.0%	.924	.963	.039	18.7%
Mean	*59.4*	*62.7*	*3.3*	*46.4%*	*.923*	*.943*	*.021*	*9.5%*

Table 2: Self-report of increased energy to engage computer

Pair	Speech Amplitude		Pen Pressure	
	Spontaneous	Prompted	Spontaneous	Prompted
1	Yes	No	No	No
2	No	No	No	No
3	No	Not sure	No	No
4	No	No	No	No
5	Yes	Yes	No	Not sure
6	No	Yes	No	Yes
7	Yes	Not sure	No	No
8	Yes	Yes	No	No
9	No	No	No	No
10	No	Yes	No	Not sure
11	Yes	Yes	No	No
12	No	Yes	No	No
Mean	*41.7%*	*50.0%*	*0%*	*8.3%*

Self-Report on Communicative Energy and Errors

As summarized in Table 2, many people were unaware that they were using greater speech amplitude when addressing the computer, and almost all were unaware of using greater pen pressure. For speech, only 41.7% of people spontaneously mentioned talking louder to the computer during three open-ended interview questions on this topic, whereas 50.0% acknowledged talking louder to the computer later 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, $\chi^2(1) = 9.25, p < .01$.

With respect to awareness of errors, 50.0% of users were not able to accurately report change in the error rate during their speech session and 50.0% could not for their pen session. In these cases they either reported an increase in errors when a decrease actually occurred, or vice versa, or they reported that errors stayed the same when more than a 10% increase or drop occurred.

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

---

the linear speech energy. Linear pen pressure was calculated using the transformation P' = .227ln(P) + .9367 – where *P* is the pen pressure value from the Anoto™ pen and P' is the linear pen pressure (force) in Newtons.

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 AND FUTURE DIRECTIONS

The present study demonstrates that people spontaneously adapt their energy level when communicating in different modalities to distinctively mark when they are addressing a computer versus human during computer-mediated meetings. During the baseline period before a user-centered energy threshold was applied, speakers' average amplitude when addressing the computer was 62.45 dB, significantly higher than 56.97 dB to a human peer. This substantial difference in amplitude that speakers use to mark computer- versus human-directed speech has been confirmed elsewhere [8]. In addition, writers' average pressure when addressing the computer was .947, again significantly higher than .923 when addressing a human partner. In short, users spontaneously adapt their energy level in both modalities to distinguish a computer versus human partner.

Whenever the computer failed to recognize that it had been addressed and to engage properly, people forcefully increased energy to clarify their intended interlocutor. In this respect, during a repair they accentuated their naturally-occurring communicative behavior by increasing their speech amplitude from an average of 59.4 dB before the miss to 62.7 dB afterwards. They likewise increased their pen pressure

from an average of .923 to .943 afterwards. These adaptations represented a substantial 46.4% and 9.5% relative increase in energy in speech amplitude and pen pressure, respectively. This discovery that people use energy, including speech amplitude and pen pressure, to mark or emphatically repair who they are addressing in a group context is distinct from past work on hyperarticulate adaptations during episodes of human-computer lexical error corrections. When an individual adult user corrects a misrecognized word by a speech recognizer, their hyperarticulate signal adaptations involve durational and articulatory effects, but not increased amplitude [12].

This research demonstrates that if a system is designed which is fundamentally adapted to users' natural communication, in this case engaging a system when a user's speech amplitude or pen pressure surpasses a user-centered threshold set for them, then they are more likely to learn and further enhance this behavior to achieve their objectives when interacting with such a system. In the present study, this further adaptation over time was most clearly observed in the speech modality, with students adapting their interpersonal speech over time to be 1.53 dB lower in amplitude. However, they also increased their amplitude by .75 dB over the session when addressing the computer. Together, these adaptations lead to an expansion of speakers' amplitude differential when addressing their computer versus human partner by an additional 2.0 dB by the session's end. Consequently, the simulated speech system showed a substantial 24.3% relative reduction in its engagement error rate by the session's end.

When using digital paper and pen, students also increased their pen pressure significantly over the course of the session from .947 to .952, accentuating their natural means of distinguishing the computer as the intended addressee. However, unlike the speech sessions, they did not significantly change their human-directed pressure or their computer-human pressure differential over the session. It is possible that these further adaptations would have occurred in the pen interface also if longer sessions and more opportunity for training were available. In addition, interactive speech is a naturally occurring interpersonal activity, and both humans and animals have evolved to use amplitude as a major cue in attracting and maintaining the attention of specific communication partners. Speech also is a public activity, and there are social prohibitions against overly loud interpersonal speech that may have encouraged the divergence observed in this study between amplitude addressed to a computer versus human partner. In contrast, non-digital pen communication among humans is a non-interactive medium. As such, it remains unclear what the origin is of users' spontaneous pen pressure differences when interacting with computer versus human interlocutors, unless it is simply a generalized behavior based on lifelong experience adapting speech communication.

During this study, students engaged the computer tutoring system over 100 times per session to obtain information as they worked collaboratively on solving math problems.

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. In fact, 7 of 12 students achieved reliabilities in the 90-100% range and 11 of 12 in the 70-100% range. 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. Using pen pressure, 10 of 12 subjects had reliabilities in the 70-100% range. In short, there was a high level of correct system engagement based exclusively on implicit cues in users' energy level during communication.

The present accuracy levels were achieved in spite of the fact that people had limited or no explicit awareness that they were using their speech amplitude or pen pressure differently to engage the computer rather than a peer. 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. Of these two, people were more aware of their amplitude than pressure, perhaps because speech is a more public and also practiced communication activity. 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.

During collaboration on complex mathematics problem solving tasks, students were able to maintain their performance level without deterioration throughout a lengthy session, in spite of engaging the system over 100 times for information and assistance. When using implicit pen pressure cues, students correctly solved 66.8% of their math problems on the first half of their session, and 72.3% on the second half. Using the implicit speech amplitude cues, they correctly solved 78.0% of the problems on the first half, and 79.8% on the second half. 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 (i.e., 70.24% for pen, versus 79.46% for speech). This performance difference only can be attributed to the interfaces students used, since the same students completed the same problems in the same orders, and the order of completing speech and pen sessions likewise was counterbalanced. This finding emphasizes the importance of future work to fortify the reliability of any pressure-based pen interface before important applications are developed. Such work could include improved approaches to developing user-centered thresholds, designing engagement techniques based on additional cues such as writing speed or height of characters, visual feedback techniques, or application of machine learning techniques. From an educational viewpoint, this finding also highlights the critical nature of designing high-quality interfaces to supporting student performance, which can be especially

fragile for lower-performing students [11].

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 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.*

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. 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. Finally, while using these implicit user-adaptive interfaces to frequently engage a tutorial system, students were able to maintain their performance at solving complex mathematics problems throughout a one-hour session. However, students' cognitive load was best managed with the speech amplitude engagement technique due to its higher reliability. 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., Lunsford, R., Wesson, R. and Oviatt, S. Prototyping novel collaborative multimodal systems: Simulation, data collection and analysis tools for the next decade. *Proc. of ICMI 2006*, ACM Press (2006), 209-216.

2. Arthur, A., Swindells, C., Oviatt, S. and Cohen, P. A High-performance dual-wizard infrastructure supporting speech and digital pen input, in submission.

3. Hinckley, K., Guimbretiere, F., Baudisch, P., Sarin, R., Agrawala, M. and Cutrell. E. The springboard: Multiple modes in one spring-loaded control, *Proc. of CHI,* ACM Press (2006), 181-190.

4. Kao. H., Hong, M. and Wah, L. Handwriting pressure: Effects of task complexity, control mode and orthographic difference, *Graphonomics: Contemporary Research in Handwriting*, ed. by H. Kao, G. van Galen & R. Hoosain, North Holland, 1986, 47-66.

5. Li, Y., Hinckley, K., Guan, Z. and Landay, J. Experimental analysis of mode switching techniques in pen-based user interfaces, *Proc. of CHI,* ACM Press (2005), 461-470.

6. 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.

7. Lunsford, R., Oviatt, S. & Coulston, R. Audio-visual cues distinguishing self- from system-directed speech in younger and older adults. *Proc. of ICMI*, ACM Press (2005), 265-272.

8. Lunsford, R., Oviatt, S. and Arthur, A. Toward open-microphone engagement for multiparty interactions, *Proc. of ICMI 2006*, ACM Press (2006), 273-280.

9. Messer, D. The identification of names in maternal speech to infants. *Journ. of Psycholinguistic Research*, *10* (1), 1981, 69-77.

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

11. Oviatt, S. L., Arthur, A. and Cohen, J. Q uiet interfaces that help students think, *Proc. of UIST,* ACM Press (2006), 191-200.

12. Oviatt, S., MacEachern, M. and Levow, G. Predicting hyperarticulate speech during human-computer error resolution, *Speech Communication* (1998), 24 (2), 1-23.

13. Paek, T., Horvitz, E., and Ringger, E. Continuous listening for unconstrained spoken dialog. *Proc. of ICSLP*, Chinese Friendship Publishers, (2000), 138-141.

14. Ramos, G., Boulos, M. and Balakrishnan, R. Pressure widgets, *Proc. of CHI*, (2004), 487-494.

15. Saund, E. and Lank, E. Stylus input and editing without prior selection of mode, *Proc. of UIST,* ACM Press (2003), 213-216.

16. Schroger, E., A neural mechanism for involuntary attention shifts to changes in auditory stimulation. *Journ. of Cognitive Neuroscience*, *8*(6), 1996, 527-539.

17. Shriberg, E., Stolcke, A. and Baron, D. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. *Proc. of Eurospeech,* (2001), 1359-1362.

18. van Turnhout, K., Terken, J., Bakx, I. and Eggen, B. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. *Proc. of ICMI*, ACM Press (2005), 175-182.