Implication of Multimodality in Ambient Interfaces

Priyamvada Tripathi
Center for Cognitive Ubiquitous Computing
School of Computing and Informatics
Arizona State University
pia@asu.edu

Sethuraman Panchanathan
Center for Cognitive Ubiquitous Computing
School of Computing and Informatics
Arizona State University
panch@asu.edu

ABSTRACT
Ambient interfaces have long held the promise of enhanced and effective human machine interaction. Ambient interfaces can adapt to human activity allowing seamless exchange of information. This goal requires a coordinated development effort that incorporates a thorough understanding of human perceptual system in the design of interfaces. In this way, ambient interfaces can not only supplement the current activities of humans but also expand their functionality to novel approaches in interaction. Since humans interact with their environments through multiple channels, multimodality is an indispensable aspect of ambient interfaces. Multimodality is a broad term that encompasses not only sensory aspects of human- machine interaction but also cognitive interaction that is responsible for a unified perception. Thus, sensory formats are essential in construction of ambient interfaces but do not constitute the complete picture. In this paper, we propose that ambience can only be achieved when multiple modalities are considered *in toto*. In addition, multimodality cannot be implemented in isolation from the desired tasks and their pertaining contexts. We propose an integrated view of multimodal interfaces that differentiates itself from the previous conception of multimodal human-machine interaction. Multimodal interfaces have become synonymous with voice-and-gesture interaction. We propose that this category of multimodality does not fully exploit or include the human user's capability of effectively interacting and communicating with the machine. Both concepts of semantic congruence and syntactic constraints can be dealt with only when we attempt to create interfaces that include the human perceptual system in its design. This can be achieved by a staged evaluation process where in each interface is associated with its joint performance value and accessibility. Besides this, the interface and human must share the same real-world model for effective reference.

Categories and Subject Descriptors
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
User Interfaces, Haptics, Multimodal Interfaces, Synergic environments, Multimedia, Modalities.

Keywords
Exclusive multimodal, Inclusive multimodal, Synergic multimodal, Accessibility.

1. INTRODUCTION
Humans are essentially multimodal. We only need to look around to discover how much we depend on our modalities working in tandem to give us information about our external world. The "do not touch" signs everywhere are symptomatic of our tendency to reach out and explore the object seen. Our modalities of vision, audio, touch, taste, and smell are always working together. Too often we complain of no flavor in food when our taste buds are working perfectly but sense of smell is not. A simple experiment can demonstrate that one modality contributes to information that we ordinarily perceive in another modality alone. When the audio perception of speech is unclear, we are greatly facilitated by the accompanying text. Our hearing also 'improves' when we can see the lips and face of the speaker while hearing him or her – the so called 'cocktail party' effect. The famous McGurk effect [20], in which human users perceive a combination of distinctively spoken and heard phonemes, is another example of our multimodal perception.

Overall, human beings are said to have the following major senses: vision, touch and kinesthesia (sometimes collectively termed as haptics), audio, balance, smell, and taste [14]. Each sense along with its physiological sensory estimate is broadly termed as modality [10]. Each modality has its own structural unimodal constraints such as spatial resolution, temporal delay, bandwidth, and nature of information that it can perceive. These are sometimes called modality-specific attributes. Apart from these, there are amodal characteristics such as magnitude, intensity, duration, and causality [27, 31].

For a long time, researchers focused on the unimodal aspects of the sense-perception. This approach enabled expansion of our understanding in each individual physiological, psychophysical, and neural organization of modalities such as in vision [26] and touch [23]. These unimodal constructs, however, ignored the interactivity each sense affords other senses for giving us unified perception and proved to be limiting in real-world scenarios which are never quite strictly confined to a single modality. A major component of human perception is what Aristotle termed as 'sense-communis' i.e. all the senses forming a unity. Sense-communis, or multimodal perception, has now been well demonstrated [4] and compels us to reconsider some classic notions about design and evaluation of human computer interaction. First, it is evident that the principles of multimodal processing are not identical to those of unimodal perception. Thus, models of unimodal interfaces do not directly translate to multimodal interfaces. Second, two or more modalities vary across a wide range in their interaction [4]. Novel approach is needed that can capture these disparate results and underlying mechanisms of multisensory integration.

All interfaces involve at least one modality for communication between human and the machine. That includes our classic computer interaction such haptic feedback in mouse and ordinary

HAS 2008, February 11-14, Quebec, Canada

DOI 10.4108/ICST.AMBISYS2008.2905

key press, auditory warning signs localizing our visual gaze, frequent visuo-motor coordination in typing, clicking, browsing, etc. Coutaz and Vaelen [6] classify such multimodal systems as 'exclusive' multimodal systems (in which only one modality is used for input or output at one time). They proposed additional category of 'synergic' multimodal systems in which two or more modalities are used concurrently for communication between human and machine. Furthermore, researchers have widely held the distinction between multimedia systems and multimodal systems by regarding the former as 'sophisticated repositories'[6] of meaningless data and associating the latter with semantic communication. Thus, Bolt's 'put that there' system [3] by this definition is a multimodal system since it combines *that* (the gesture input) with *there* (speech input) into a single command for the interface. Since Bolt, multimodal interfaces have become synonymous with the voice-and-gesture interfaces.

The original goal of multimodal interfaces thus was to have machine behave like a human and seamless interaction meant we can talk to machine and instruct without ever feeling the loss of a human. In this conception, there are four levels of interaction [27]: (1) the low-level or signal-level interaction, (2) the information level which concerns bandwidth and digital communication (3) the cognitive level, which pertains to the pattern recognition techniques and 'top-down' modeling employed for disambiguation or fusion (4) The intention or the goal level that focuses on the goal-level behavior such as I want to find the price of a watch vs. I want to click on a webpage link. We propose that this categorization, while appearing expansive at first glance, may be somewhat limiting in the conception of multimodal interfaces. First, this categorization includes multimodal interaction at the lowest level of interaction which is directly opposite to the current view of multimodality in humans. This categorization also requires explicit identification of what the user wants and mapping it to the exact specifications of what interface can do.

The 'semantic gap' and hierarchical view of human perception and interface design will never allow us to leverage human perceptual system for maximal functionality and ubiquitous interaction. First, we must differentiate between multimodal interfaces and 'communicative' interfaces. Multimodal interfaces are truly interactive or 'coaptive' interfaces that complement the human perceptual system in toto. Communication or exchange of information between user and the machine is, no doubt, an aspect of the multimodal interfaces but it does not provide the complete picture. In addition to the aspects of communication, multimodal interfaces by their definition include factors such as context, level of mental effort the interface demands, and mutual compatibility of the tasks and modalities that concurrently occupy the interface design.

In order to achieve this goal, we need to design interfaces that are geared to complement the user. Figure 1 shows our conception of ambient interfaces that harmonize with the human user. We divide the perception of the human user into two aspects: (1) Sensory, (2) Cognitive. The corresponding aspects of the machine are termed are called (1) Format, (2) Process. Format/Sensory corresponds to the modality specific constraints or structural limitations of the human perception. The sensory aspects of the user have always had to agree with format of the machine to give a more naturalistic experience. Not until recently, Charlie Chaplin movies, even though entertaining, were partly humorous for their disagreement with human visual perception. Even to present day, we strive for the congruence of format of presentation with human eye and removing artifacts and pixelation that are still apparent in big screen televisions. In recent years, with addition of more modalities such as audio and haptics, similar limitations have presented themselves in these extended interfaces. Thus, format congruency is essential for ambience in multimodal interfaces. Beyond the unimodal aspects, we have proposed the term 'cognition'. Cognition, is broad term, and includes several aspects of the human perceptual system. Current research reflects that there no definite boundaries of perception and cognition in human brain. At best, we can visualize our brain as a set of interacting nodes, each with some specialized function and assigned a certain weight according to the given context, goal, and user's mental states (like fatigue, alertness, distraction level, etc.). These factors are highly contextual and must be accommodated in the design evaluation of an interface. The last aspect we consider important is that both interface and human need to share the same real-world referent. For example, if the interface deals with spatial attributes, then human users must be able to relate it spatially in their egocentric frame. This can either be trained or translated directly from real-world conceptions. This is consistent with Sutherland's [30] dream of the "ultimate display". According to him, "The ultimate display would, of course, be a room within which the computer can control the existence of matter. A chair displayed in such a room would be good enough to sit in. Handcuffs displayed in such a room would be confining, and a bullet displayed in such a room would be fatal. With appropriate programming such a display could literally be the Wonderland into which Alice walked."

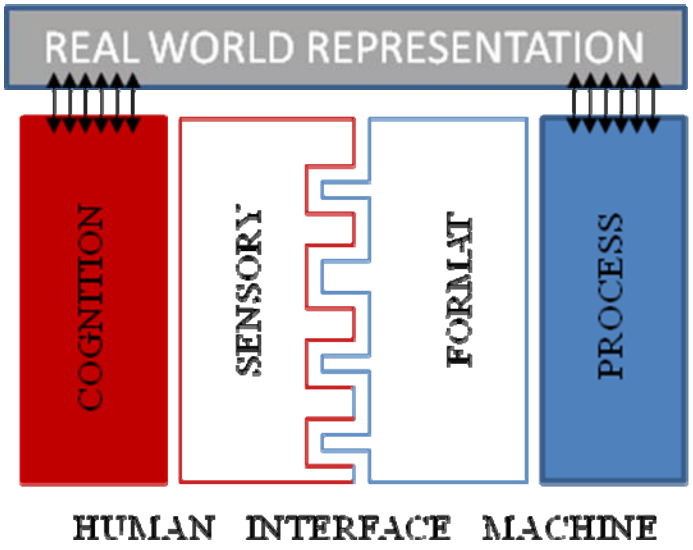

Figure 1. Ambient Human Computer Interface can be visualized as a grooved parts from human and machine that fit together to give a seamless unified perception.

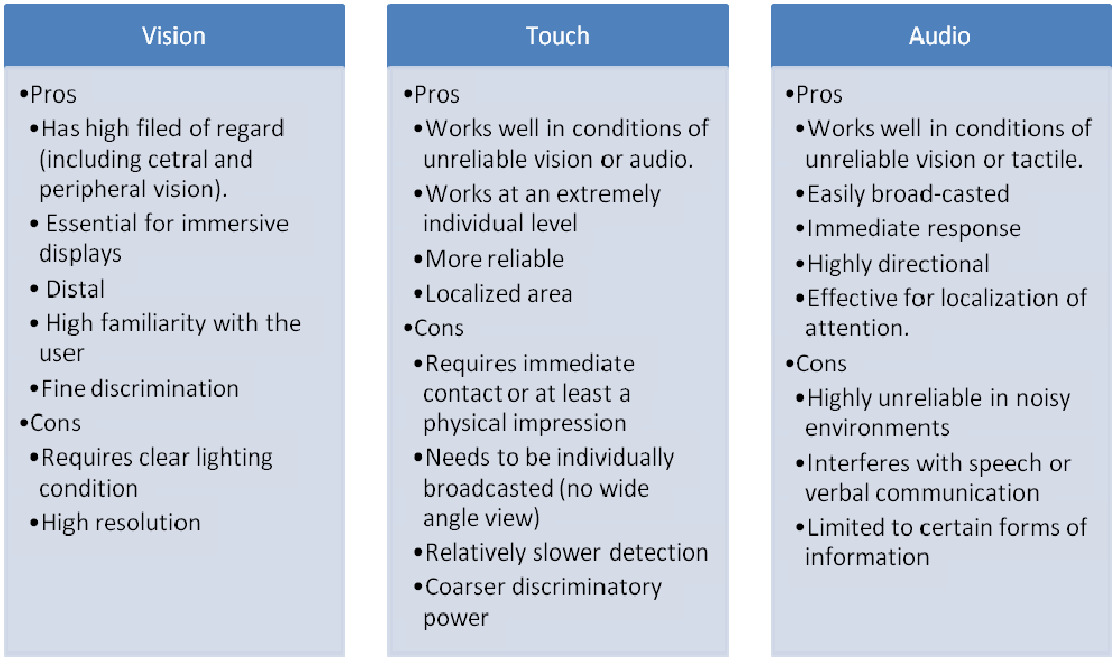

Figure 2. Pros and Cons of each modality for user interfaces.

2. INTERNAL FACTORS IN MULTIMODAL AMBIENT SYSTEMS

Humans have five major modalities each of which is designed to acquire information about the external world in some special form. Vision is specially designed for color, luminance, audio for loudness and pitch, haptics modality which constitutes tactile for shape, pressure, temperature, pain, and kinesthesia for movement and motor coordination. Gustatory and Olfactory sense are relatively underutilized in the context of machine environments and not discussed here. It should be pointed out, however, that these other senses play an important role in nature for warnings, territorial marking and hold potential for certain specialized purposes as well as general alertness and awareness. In Figure 2, we list the possible advantages and disadvantages for vision, audio, and haptics in user interfaces (see Stanney et al. [29] for detailed guidelines). One should note that some limitations of the most intuitive modality (Lederman [14] termed this as the ecologically validity of modality; Welch and Warren called it the 'modality-appropriateness of a task [30]) for the given purpose can be circumvented by *recoding* a task to another modality. For example, a warning signal can be given through vision, audio, and/or touch. The constraints of reaction time or environment (high noise or low illumination) may further determine the selection of modality in each case.

Most interfaces, however, are highly complex in nature. In such conditions, one cannot ignore the effect of information presentation in one modality to another. In some cases, it may be even advantageous to consider the effect in order to exploit the human perceptual system for effective information presentation. Empirical results now show unequivocally that several sensory estimates that were earlier assumed to be unrelated have commonalities. Marks [19] referred to these as 'communalities' and categorized them as follows: (1) Informational Communalities: Different sensory systems provide common or equivalent information such as size or form or duration, (2) Psychophysical Communalities: Functional similarities in the ways in which sensations and perceptions depend on stimulus parameters as intensity, qualitative structure, and distribution in space and time, (3) Phenomenological Communalities: Similarities directly perceived between qualities of perceptual experiences in different modalities (e.g., Synaesthesia).

These interactions between modalities are now broadly considered with respect to the interaction of the individual 'cues' in each modality. A 'cue' can be defined as any sensory information that gives rise to a sensory estimate [10]. Interaction of these cues are broadly categorized into four categories [7, 16]. Let s1 and s2 be two independent cues either within modality or across distinct modalities and I is the information carried by a cue, then:

a) Redundant cues ($I(s1) \approx I(s2)$): Information carried is the same

b) Concordant cues (variance (s1)< variance(s2) or RT(s1) <RT(s2)): One modality is more reliable or faster over another

c) Discrepant cues ($I(s1) \neq I(s2)$ and $s1 = s2$): They conflict in their information

d) Complementary cues ($I(s1) \cap$ intersection $I(s2) \neq$ NULL and $s1 \neq s2$): Qualitatively different but carry helping each-other's content.

These cues can result in various forms of interaction. For highly discrepant cues, *vetoing* is observed in which one modality completely dominates final perception. Other forms of interaction include mutual *disambiguation* in which cues contribute to clarifying another cue, *cooperation* wherein cues combine together for final percept, and *promotion* which is similar to mutual disambiguation [5, 12]. Thus, the affect of one modality on another can be either quantitative, such as reaction time or accuracy, or qualitative such as improved perception. These interactions are known under various names such as inter-sensory facilitation, priming, intermodal transfer, cross-modal attention, multisensory integration, each focusing on one individual aspect of the interaction.

Models of multimodal integration can be weakly coupled or strongly coupled [5]. In weakly coupled models, output of each modality is independent of the outputs from other modalities and is combined in the multimodal module while in strongly coupled models, output of one modality may influence the output of another modality. A common example for the weakly coupled model is found in texture discrimination by touch and vision wherein the final perceptual value if intermediate value of that perceived in that in vision and that in touch [15].On the other hand, some models in depth perception support the strongly coupled hypothesis [5]. Figure 3 illustrates some common example of models of multisensory integration.

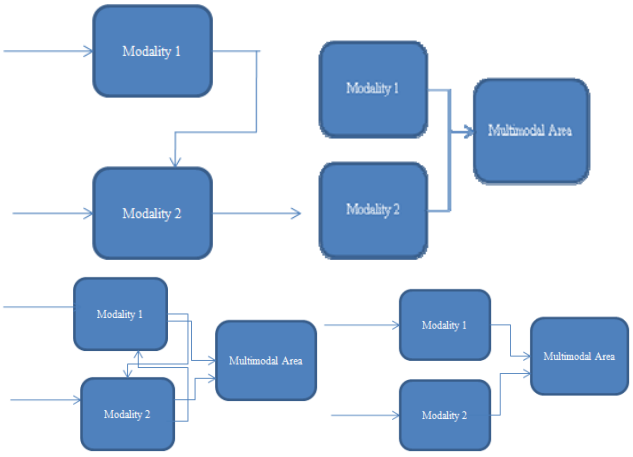

Figure 3. Several possibilities exist in Multisensory Integration.

Most researchers, however, agree that given spatio-temporal concurrency of inputs in two or more modalities, the multimodal response is a super additive function of the unimodal outputs. For example, Benoit et al. [2] tested recognition accuracy with audio alone (unimodal condition), and audio and vision (multimodal condition) together. They found that in clear conditions, the multimodal performance is close to the unimodal performance but in degraded acoustic condition, multimodal listeners could understand 12 out of 18 items presented while unimodal subjects could understand none at all. In 2002, Ernst and Banks [9] demonstrated that when noise is added to one modality (in their case vision), users increasingly rely on other modality (touch) that is delivering the same information. Backus et al. [1] found similar effects in stereo perception. Thus, these models can be used in conjunction with unimodal constraints or environmental factors to get optimal behavioral results from human users.

The knowledge of multimodal integration is incomplete without the three essential criteria that enable such interaction. These are commonly [21, 27] listed as: (1) Temporal coherency, (2) Spatial coherency. In addition, we propose a third criteria: feature coherency [8, 22]. Typically it is required that when information needs to be perceived by two or more modalities simultaneously, only a certain lag between the two streams can be tolerated. Since each modality has its own specific processing time range, careful measures are required for synchronized delivery of sense-data.

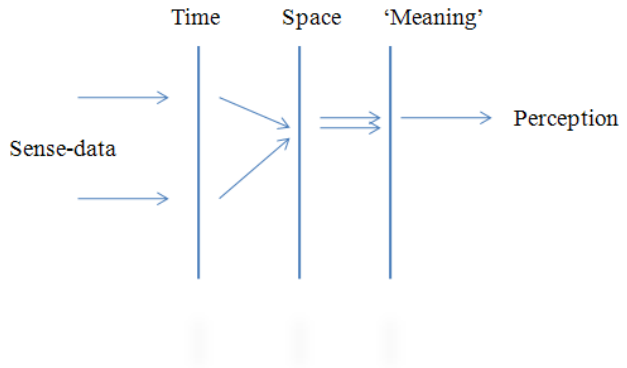

Figure 4. Temporal coherency, spatial coherency, and semantic congruence are three essential steps for achieving multimodal perception.

Empirical Research [28] has pointed out temporal synchrony of input results in a response enhancement while temporal asynchrony results in response suppression (sometimes known as inverse effectiveness). Spatial synchronization requires delivery of information to the sensory system so that we can perceive the semantically congruent information to have come from the same location. Since this perception is usually dictated by the dominance of one content or modality over another (such as that in a ventriloquist), there seems to a large flexibility in this category for design of interfaces. However, unwanted effects will be observed as that of confusion or distraction if spatial localization is not observed. Feature synchronization may more appropriately called semantic congruence since it deals with the higher order information. If the object or environmental information delivered via two or more modalities agrees with each-other, user will respond effectively and quickly, but varied responses may be seen if this is not so. Some users may immediately perceive conflict and ambience will suffer dramatically.

3. EXTERNAL FACTORS IN MULTIMODAL AMBIENT INTERFACES

Context is an important factor in determining the meaning of the stimulus presented. Context contributes to the minimizing the ambiguity of the environmental stimuli, defining the relative difficulty of the task that is to be performed. Figure 5 gives a common example of contextual removal of ambiguity in perception. The top row in the figure gives an ambiguous perception of the symbol 7 but the lower row makes it immediately clear that the symbol is the number 7. Context is, hence, concomitant information that is available to the user during interaction. In this figure, we have two independent contexts, one in top row, and one in the bottom row, that determine the

availability of the meaning or level of clarity in the given task (that of identifying letter 7). Thus, we define an associated concept with context which we called accessibility.

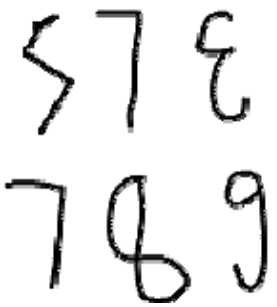

Figure 5. Top row letter 7 is unclear but bottom row the context of numbers removes this ambiguity immediately.

Higgins [11] originally defined accessibility as "the ease with which particular mental contents come to mind". Daniel Kahneman, in his Nobel Prize lecture in 2002, expressed is as a continuum of 'cognitive fluency' in perception. On one end, he said, there are operations that are effortless and rapid, and on the other hand, those that require significant mental effort to process the stimuli. Sometimes mere representation of the task can place is on the one end or other. This can be illustrated using the visual search results from Treisman and Gelade [26]. In their experiments, they showed a display to the participants which has two types of objects. The goal of the participant was to identify and select the target stimuli from the display of embedded stimuli of a slightly different type. They found that higher the number of dimensions (or features) the target differed from that distracter stimuli, the more time it took for the participants to find the target. We can say that these targets, with increasing complexity, were on the higher end of the accessibility continuum.

Accessibility can also be dependent on the nature of coding or the modality of presentation. For example, we may find it hard to guess the depth from visual estimate alone but allowing tactile modality can significantly allow quicker perception of this information. A task such as comparison of size, shape, orientation, or calculations of perceptual dimensions such as height or width of the object can be reduced by allowing more modalities to access the stimuli.

Accessibility is not restricted to sensory dimensions alone. Lederman et al. [17] in their multidimensional scaling study on vision and haptic perception of objects showed that depending on the instructions, users either categorized the objects based on haptic features or visual features. Although accessibility is largely a bottom-up concept, there are clear examples of manipulating the bias of the user towards one direction more than the other. In the context of user interfaces, therefore, one needs to accommodate accessibility as a function of mental effort needed to perform a task. In this sense, if we have two tasks one of which requires more effort than another, than quite automatically, ordering of the tasks will take place in correlation with their dimension of accessibility.

Accessibility of a task in a user interface thus relies on two main criteria (1) how long does it take to complete the given task (2) how much does it interfere with another task with which it is being performed. Accessibility is, therefore, indicative of mental effort that is needed to process a task and is closely interlinked with the context in which a task is being performed. Accessibility can be enhanced by subjective awareness, practice, or explicit instructions to the user and encompasses varied concepts such as saliency and preferential allocation of attention. Notice that we resist the temptation to describe accessibility in terms of mental resources. Even though resource theory has been widely successful in explaining some empirical results [31], due to wide conflicts in results across literature, and relative flexibility in adding or deleting resources to accommodate these conflicting results, we believe that resource view will prove to be too precarious for designing multimodal interfaces (see [18] for criticisms of the resource view).

4. COAPTIVE MULTIMODAL INTERFACES

A task-centric view is different from both resource-centric view (where resources are assumed to be limited) and a data-centric view (where sufficiency of information available is a criterion for accuracy) [18]. While resource centric view emphasizes the restricted capacity of human cognition to process multiple tasks simultaneously, particularly when they share the same resources, it does not address the issue of the suitability of a task, or information format, with respect to presentation modalities. The nature of interaction of multiple tasks is known to vary with respect to different modalities. The McGurk effect [20] is an example of interaction of differing information content received in separate modalities. Thus, task-centric view enables identification of the format of the tasks presented to each modality and employment of the nature of their interactions as an effective tool for multimodal environments. We essentially a task centric view to interfaces in which we describe each interface as a combination of tasks and modalities with their corresponding performances. This can be represented as a task-modality matrix (see Figure 6).

	A	**B**	**..**	**Z**	**Modality**
1	(1,A)	(1,B)	..	(1,Z)	
2	(2,A)	(2,B)	..	(2,Z)	
:	:	:	..		
n	(n, A)	(n, B)	..	(n, Z)	
Task					

Figure 6. Task-modality Matrix

A task is defined as any activity performed by the user with an intended goal. For example, a user can be alerted by an audio tone (such as a beep), a text dialogue, or both. In the first case, task is warning and modality is audio, in the next case, task is warning but the modality is now vision, and in the last case, warning is simultaneously conveyed to both audio and vision (a multimodal task). Occasionally same modality may be concurrently performing two tasks. For example, we may use audio for

MODALITY x TASK		PERFORMANCE							
		Task 1		Task 2		…	Task n		
		Metrics	S.D.	Metrics	S.D.		Metrics	S.D.
	(1,A)								
stage 1	(1,B)								
	(2,A)								
	(B,2)								
	…								
	(1,A)(1,B)								
stage 2	(2,A)(1,B)								
	(1,A)(2,B)								
	(2,A)(2,B)								
	…								
	(1,A)(1,B)(2,B)								
stage 3	(1,A)(2,A)(2,B)								
	…								

Figure 7 . Framework for evaluation of Task –modality interference.

perceiving warning and also be listening to music at the same time. At the time of concurrent performance, both tasks will affect each-other and result in joint performance. In these concurrent occurrences, our performance in either task alone is of little consequence to our joint performance. We can treat these unimodal contexts as control levels for comparison of multimodal contexts.

Most interfaces, therefore, can be represented as a combination of tasks and modalities. We propose that the following factors that need to be kept in mind when designing multimodal ambient interfaces:

1) What modalities are suitable for the given task(s)
2) The performance metrics of a task(s) within each modality (unimodal context),
3) How will the nature of task(s) change in presence of another modality or task (multimodal context)

We now propose a staged process that can be followed wherein we move from 'exclusive' to 'coaptive' or ambient environments. The framework thus proposed only includes standard deviation and performance metric. The use of performance metric and standard deviation lends itself suitable to several psychological tests. To begin with, analysis of variance can be run between different factors (modalities) or different levels (contexts) for testing significance of interaction. If no interaction is found, we can easily put that particular combination together without much hindrance to the overall performance. Using empirical insights from psychology [9], we can say that if the new standard deviation is lower than the each standard deviation in previous condition, the combination is complementary. However, if it is more, combination is contradictory, and if it neutral, combination may either be redundant or neutral. A combined consideration of all metrics may be employed to use a certain task-modality combination. For example, if a task combination has neutral influence in performance accuracy but lowers standard deviation, it should be preferred. This can be generalized intuitively to other conditions.

The steps to the evaluation and construction of multimodal systems are listed as follows (refer to Figure 7):

1) Select intended tasks of the interface
2) Map each tasks to possible modalities (one task may be recoded on a non-intuitive modality)
3) Pick the combination from task grid and place it in the stage 1, stage 2, etc. shown in Figure 7.
4) Select user group for evaluation
5) Divide the user group into different contexts, if context is desirable factor for evaluation.
6) Measure the performance of each user in the given condition for m trials.
7) Calculate mean performance and standard deviation of group (if subgroup evaluation is preferred, select subgroups)
8) Run statistical analysis
9) If no significant decrease in performance and/or increase in standard deviation is observed in some combination over other, pick this combination.
10) Check whether the combination meets the criteria of optimality.

Thus, we have a generic methodology that can be used for designing and evaluation for multimodal interfaces.

Around the concept of resource, Norman and Bobrow [24, 25] proposed performance operating characteristics curve (POC) to visualize the difficulty of a pair of tasks. POCs plot the performance of one task with respect to performance of another task. For example, if we have task A and task B being

performed together, and performance of task A is as accurate as it was in when performed alone (control condition) while task B is also unaffected, then task A and task B are said to have no mutual interference. However, let us say if there is a task C, which when performed with task A reduces task A's performance and itself deteriorates to a lower performance level then task A and task C are said to be highly interfering. Thus, performance A and performance B are plotted in two dimensions. We extend a similar notion to multimodal interfaces. Multimodal scenarios (interfaces) can be visualized in a three dimensional space where three axis are tasks, performances, and modalities. In Figure 8, we have the following combination: Task 1(Modality A), Task 2 (Modality B), Task 3(Modality A) and three joint performances corresponding to the tasks. If we get a leveled plane, we can say that each task in the interface is equally accessible. This slope reflects the tilt of the plane which could be negative, positive, or neutral. Since any combination of tasks and modalities is possible, the similar concept of the slope can be extended to various geometric shapes in the given coordinate space.

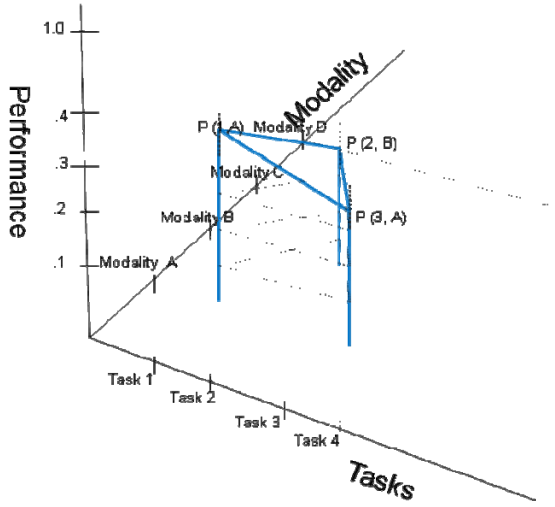

Figure 8. Planar representation of task interactions in the design of multimodal interfaces

5. EXAMPLE SCENARIO: QUANTITATIVE

A generic solution to multimodal ambient system is desirable but may not be possible in near future. This is also complicated by the fact that interfaces are highly customized environments for specific applications or group of users. We propose a novel methodological framework that be employed to assess the coherency of task grid and context. First, we need to define the set of goals that need to be accomplished. Then, we assign these goals in form of tasks to certain modalities. Our first goal is target the intuitive modalities and check whether combination yields optimum performance. If not, we reassign the tasks to new modalities in new formats and iteratively assess the joint performance of the users until optimal combination is reached. We can approach this by assessing performance in stages. Stage 1 is one-to-one stage (or unimodal stage). Stage 2 onwards we assess the subsequent combinations.

This methodology is illustrated by the following empirical study conducted to evaluate the combination of two tasks: (1) Mediated Object Recognition (2) Verbal Comprehension. Mediated object recognition is a semantic representation of the object categories that is communicated to the user via a sequence of vibro-tactile codes on hands or transcribed audio message sequences. This is shown in Figure 9.

Tactile object recognition was proposed by Kahol et al. [13] and involves sending vibrotactile cues of two durations (short, long) with same intensity. Information about the overall shape, texture, material, size, height, width of the object are categorized into perceptually salient classes such as rough, medium and smooth for texture. Each class is transcribed by the means the two vibratory patterns and sent to specific parts of the hand through a

	Touch (A)	Audio (B)	Modality
Object Recognition (1)	x	x	
Comprehension (2)	-	x	
Task			

Figure 9. An example multimodal scenario

specially designed glove with six tactile motors. For example, overall shape information (class of the object) is sent to the palm of the user. The objects that were used are divided into four classes (cup, bowl, glass, and irregular) and each class varied along four dimensions (texture, shape, size, and material). The object data set consisted of 106 objects that belonged to one of these unique combinations. The subjects were trained to discriminate between objects based on one or more features. In the case of auditory cues, verbal labels of the perceptual classes are presented to the user through a synthesized voice describing the object. Reading comprehension task consisted of paragraphs similar to those in a beginner level comprehension tasks in English as a Second Language (ESL) tests.

5.1.1 Participants

8 individuals (age range: 20 to 40 years; normal hearing and tactile abilities; all right handed; 4 females and 4 males) participated. All subjects were fluent in American English. A written consent was obtained prior to the study and a compensation of $40 was provided to all participants. The study was approved by the university institutional review board.

5.1.2 Procedure

Subjects were made to sit in front of a set of speakers with one piece monotonic head phones. They also wore Immersion data glove® with six vibrotactile actuators. Tactile vibrations giving characteristics of the object such as shape, size, texture and material were passed through the gloves. Audio cues delivered object features through a verbal spoken stream presented through the one piece headphone. Comprehension tasks were played out from ambient speakers placed in front of the subject.

	MODALITY x TASK	PERFORMANCE			
		Task 1		Task 2	
		Accuracy	S.D.	Accuracy	S.D.
stage 1	(1,A)	96.875	1.38873015	-	-
	(1,B)	98.96	0.353553391	-	-
	(2,B)	-	-	91.67	1.069044968
stage 2	(1,A)(2,B)	96.88	0.51754917	81.25	1.457737974
	(1,B)(2,B)	85.42	1.488047618	69.79	1.407885953

Figure 10. Results from the quantitative scenario

In the experiment, subjects received passages for reading comprehension. At randomly selected time instant during the reading comprehension task, object features were delivered through haptic cues or auditory cues. Subjects performed yes/no tasks in a within subject design (all subjects performed all conditions). For cueing tasks, the object information that was cued was tested for recognition after the tasks. Comprehension tasks were followed by True/False question. After completion of each cueing and comprehension set, participants were asked three yes/no questions in each object recognition and reading comprehension. For object recognition, we handed three objects one by one and asked participants to respond with 'yes' if the object was the object that was cued for and 'no' if it was not. Participants were told that none of the three objects may be correct or more than one object may correspond to the cueing. This ensured that each answer was independent of other answers in a given trial. The experiment was conducted by two experimenters alternately to remove any bias by the experimenter. None of the participants were informed about the cueing methodology prior to the training. Participants were recruited from outside the university to ensure their neutrality. Each participant came for four sessions, first three sessions being training sessions for audio and tactile cueing. All participants had achieved 100% accuracy for each audio and tactile cueing at the end of the third training session. The collection of data was randomized across the subjects and conditions in a counterbalanced fashion.

5.1.3 Results

Figure 10 shows the results of this scenario. First, stage 1 performance was recorded. In stage 2, the new task was added and tested. Thus, (1, A)(2,B) records performance of subjects in comprehension and tactile object recognition. (1, B)(2, B) records performance of subjects in comprehension and audio object recognition. In these results, we can see that tactile object recognition actually saw an increase in performance with audio object recognition suffered a decrease with comprehension task. Comprehension performance was lower in both combination but decrement was higher with audio object recognition. Notice that we have also recorded standard deviation in the table. At a quick glance, we notice that standard deviation of performance in comprehension is increased by both audio and tactile object recognition by about the same amount. However, standard deviation of tactile object recognition is greatly reduced while audio object recognition is greatly increases (almost by 66%).

Audio object recognition also suffers from great decrement in its performance and is accompanied by significant decrease in the performance of comprehension task. This is probably so because both audio and speech comprehension task share the same modality. If the need be, comprehension can be done by other modalities such as reading comprehension (visual) and performance can further be tested. In this scenario, we considered it sufficient since the goal was to find which of the two modalities touch or audio may be more suitable for object recognition. Also, participants repeatedly reported higher level of stress in audio object recognition when combined with reading comprehension. This may be reflected in the increased standard deviations of the tasks.

6. EXAMPLE SCENARIO: QUALITATIVE

In order to demonstrate a possible qualitative interaction, we evaluated data from Lederman and Abbott [15] and fitted it in our framework. Lederman and Abbott gave individual texture (grit) values to touch, vision, and touch and vision and found that estimate in touch and vision was an intermediate value of that in touch and that of vision. The scenario is shown is Figure 11. Here the grit value of the standard given in touch alone was 60.0 grit value for vision alone was 150.0. Subjects did a match-to-sample task in a between-subject design. Results are in Figure 12. If we notice the standard deviation, we can see that standard deviation of final performance is lesser that of tactual texture perception and not very different in the case of visual tactual perception while the mutual bias is about 50% from both visual and tactual tasks. Thus, in this case, the two modalities act in complementary manner to produce the final perceptual result.

	Touch (A)	Vision (B)	Modality
Texture perception (1)	60	150	
Task			

Figure 11. Scenario from Lederman and Abbott [15]

7. CONCLUSION

The construction of multimodal ambient systems depends on technological advances as well as psychological literature. A

MODALITY x TASK		PERFORMANCE	
		Task 1	
		Texture Estimation	S. D.
stage 1	(1,A)	72.33	42.8
	(1, B)	114.33	33.9
stage 2	(1,A)(1,B)	92.67	31.6

Figure 12. Results from Lederman and Abbott [14]

thorough understanding of human perception will not only impel the development of the right tools and techniques, it will also contribute to expanded functionality and enhanced interaction between humans and machines. For a naturalistic interaction, it is important that interfaces are developed with human-in-the-loop. We propose that multimodal systems must be conceptualized in toto. Current human computer interaction largely assumes a neutral effect between the various tasks across modalities of the human operator. This assumption of independence in multimodal environment fails when we consider studies that have pointed out that two or more sensory modalities may influence each-other in variety of ways. No unimodal model will directly transfer to multimodal context and hence, will only go so far in development of ubiquitous environments. The framework proposed in this paper enables customized approach to multisensory user interfaces and relies purely on empirical data.

8. REFERENCES

[1] Backus, B.T., Banks, M.S., van Ee, R. and Crowell, J.A. Horizontal and vertical disparity, eye position, and stereoscopic slant perception Vision Research, 39 (6). 1143-1170.

[2] Benoit, C., Mohamadi, T. and Kandel, S. Effects of phoentic context on audio-visual intelligibility of french speech in noise. Journal of Speech and Hearing Research, 37. 1195-1203.

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

[4] Calvert, G., Spence, C. and Stein, B.E. (eds.). The handbook of multisensory processes. MIT Press, Cambridge, Mass., 2004.

[5] Clark, J.J. and Yuille, A.L. Data fusion for sensory information processing systems. Kluwer Academic Publishers, Boston, 1990.

[6] Coutaz, J. and Caelen, J., A Taxonomy for Multimedia and Multimodal User Interfaces. in 1st ERCIM Workshop on Multimodal Human-Computer Interaction, (Lisbon, 1991).

[7] Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J. and Young, R.M., Four easy pieces for assessing the usability of multimodal interaction: the CARE properties. in INTERACT 1995, (1995), 115-120.

[8] Driver, J. and Spence, C. Multisensory perception: Beyond modularity and convergence. Current Biology, 10. R731-R735.

[9] Ernst, M.O. and Banks, M.S. Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415. 429-433.

[10] Ernst, M.O. and Bu¨lthoff, H.H. Merging the senses into a robust percept. Trends in Cognitive Sciences, 8. 162-169.

[11] Higgins, E.T. Knowledge activation: Accessibility, applicability, and salience. in Higgins, E.T. and Kruglanski, A. eds. Social psychology: Handbook of basic principles, Guilford Press, New York, 1996, 133-168.

[12] Johnston, E.B., Cumming, B.G. and J., P.A. Integration of depth modules: Stereopsis and texture. Vision Research, 33. 813-826.

[13] Kahol, K., Tripathi, P., McDaniel, T., Bratton, L. and Panchanathan, S. Modeling context in haptic perception, rendering, and visualization. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 2 (3). 219-240.

[14] Kandel, E.R., Schwartz, J.H. and Jessell, T.M. Principles of Neural Science. McGraw-Hill 2000.

[15] Lederman, S.J. and Abbott, S.G. Texture Perception: Studies of intersensory organization using a discrepancy paradigm and visual vs tactual psychophysics. Journal of Experimental Psychology: Human perception and performance, 7. 902-915.

[16] Lederman, S.J. and Klatzky, R.L. Multisensory texture perception in Calvert, G., Spence, C. and Stein, B.E. eds. The handbook of multisensory processes, MIT Press, Cambridge, Mass., 2004, 107-122.

[17] Lederman, S.J., Thorne, G. and Jones, B. Perception of texture by vision and touch: multidimensionality and intersensory integration. Journal of Experimental Psychology : Human Perception and Performance, 12 (2). 169-180.

[18] Luck, S.J. and Vecera, S.P. Attention. in Yantis, S. ed. Steven's Handbook of Experimental Psychology: Sensation and Perception, 2002, 235-286.

[19] Marks, L.E. On cross-modal similarity: auditory-visual interactions in speeded discrimination. Journal of Experimental Psychology: Humam Perception and Performance, 13 (3). 384-394.

[20] McGurk, H. and MacDonald, J. Hearing lips and seeing voices. Nature, 264. 746-748. .

[21] Meredith, M.A. On the neuronal basis for multisensory convergence: a brief overview Cognitive Brain Research, 14 (10). 31-40.

[22] Meredith, M.A., Nemitz , J.W. and Stein, B.E. Determinants of Multisensory Integration in Superior Colliculus.I. Temporal factors. Journal of Neuroscience, 7 (10). 3215-3229.

[23] Mountcastle, V.B. The Sensory Hand: Neural Mechanisms of Somatic Sensation. Harvard University Press, Cambridge, MA, 2005.

[24] Navon, D. and Gopher, D. On the Economy of the Human-Processing System. Psychological Review, 86 (3). 214-255.

[25] Norman, D.A. and Bobrow, D.G. On the Analysis of Performance Operating Characteristics. Psychological Review, 83 (6). 508-510.

[26] Rolls, E.T. and Deco, G. Computational Neuroscience of Vision. Oxford University Press, 2002.

[27] Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoit, C., Guiard-Marigny, T., LeGoff, B., Robert-Ribes, J., Adjoudani, A., Defee, I., Munch, S., Hartung, K. and Blauert, J. A taxonomy ofmultimodal interaction in the human information processing system, Esprit Project MIAMI, 1995.

[28] Spence, C. and Driver, J. Crossmodal space and crossmodal attention. Oxford University Press, Oxford ; New York, 2004.

[29] Stanney, K.S., Shatha; Reeves, Leah; Hale, Kelly; Buff, Wendi; Bowers, Clint; Goldiez, Brian; Nicholson, Denise; Lackey, Stephanie A Paradigm Shift in Interactive Computing: Deriving Multimodal Design Principles from Behavioral and Neurological Foundations International Journal of Human-Computer Interaction, 17 (2). 229-257.

[30] Sutherland, I.E., The Ultimate Display. in IFIPS Congress, (1965), 506-508.

[31] Welch, R.B. Meaning, attention and the "Unity Assumption" in the intersensory bias of spatial and temporal perceptions. in Aschersleben, G., Bachmann, T. and Müsseler, J. eds. Advances in psychology: Cognitive contributions to the perception of spatial and temporal events, Elsevier, 1999, 371-387.

[32] Wickens, C.D. Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science, 3(2). 159-177.